# AN INTRODUCTION TO STATISTICAL METHODS

A TEXTBOOK FOR COLLEGE STUDENTS
A MANUAL FOR STATISTICIANS
AND BUSINESS EXECUTIVES

BY

HORACE SECRIST, PH.D.
PROFESSOR OF ECONOMICS AND STATISTICS AND DIRECTOR, THE BUREAU OF BUSINESS RESEARCH, NORTHWESTERN UNIVERSITY
Author: *Readings and Problems in Statistical Methods*, Etc.

AGRICULTURAL RESEARCH INSTITUTE
No. 1528
11 APR 1927
COIMBATORE

REVISED EDITION
(Entirely Rewritten and Enlarged)

New York
THE MACMILLAN COMPANY
1925

*All rights reserved*

THE MACMILLAN COMPANY
NEW YORK · BOSTON · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO

MACMILLAN & CO., LIMITED
LONDON · BOMBAY · CALCUTTA
MELBOURNE

THE MACMILLAN CO. OF CANADA, LTD.
TORONTO

Copyright, 1917 and 1925,
By THE MACMILLAN COMPANY.

Set up and electrotyped. Published December, 1917.
Revised edition, June, 1925.

PRINTED IN THE UNITED STATES OF AMERICA BY
THE BERWICK & SMITH CO.

To

The Memory of

JOHN FILLMORE HAYFORD,

MY FRIEND AND COUNSELOR

# PREFACE TO REVISED EDITION

During the eight years since the first edition of this book appeared, there has been a remarkable development in the use of statistics and statistical methods. This has come about in part because of the need for quantitative data during and following the World War, and also because of the growing appreciation that social, political, business, and economic policies should rest upon a factual basis.

The development has taken a variety of forms. Statistics and statistical methods now constitute an important part of college and university instruction; banks, research agencies, and the government, particularly, publish statistics on a wide variety of topics relating to trade and industry, social and industrial progress, and business conditions. Moreover, the larger business firms now have their own statistical departments in which they collect and interpret facts about their own affairs, and in which they use those collected by others. There is scarcely an economic or social issue which is not being treated statistically. A renaissance of interest in all phases of statistics seems to have captivated the business and social world.

While this is gratifying, it raises two questions in which teachers of statistics and practicing statisticians are vitally interested: (1) What type of training is necessary in order to develop men and women skilled in the preparation, use, and interpretation of statistics? and (2) How should the introductory subject matter of statistics and statistical methods be presented? The writer, during the past fifteen years, has given the better part of his time and attention to a consideration of these and similar inquiries, and the revised edition of this book contains his answers.

This edition, while retaining the distinctive features of the one which it supplants, records the progress which has been made in the technique and use of statistics since 1917. The subject matter is discussed in keeping with the well-established pedagogical principle that skill and judgment in the use of statistics can be best acquired when the methods are presented in the order in which they are used in statistical analysis.

The book, it is hoped, is more than a "statistical arithmetic," or even a compendium of statistical practices. A conscious effort has been made to give it body and substance, and to state and illustrate the principles back of numerical calculation and manipulation. Mathematical formulæ and descriptive methods of how to use statistics, while fully explained, are discussed in connection with the logical place which they hold in scientific thinking. Statistical analysis, requiring as it does observation of facts, their measurement, suitable analysis, and logical inference is treated broadly and fundamentally. The book is concerned with the statistical ways in which each of the steps in constructive thinking should be carried out. It is intended to be an essay in applied logic. While designed as an introduction to the subject, it is broad enough in scope, it is believed, to supply the basis for a thorough understanding of the elementary principles of statistics and statistical methods.

In the revision, the book has been entirely rewritten, enlarged, simplified, rearranged, more fully illustrated, and, it is hoped, the principles more accurately stated. Among the changes that have been made are the following: Chapters II and XI, in the old book, are now Chapters II and III, and X and XII, respectively. New chapters on *The Theory of Probability and some Properties of the Normal Law of Error Distribution,* and on *The Treatment and Correlation of Time Series* have been added. Those relating to *The Principles of Index Number Making and Using,* and *American Index Numbers Described and Compared,* have been entirely recast and

given new positions in the order of treatment. Both the principles and methods of constructing index numbers of quantities, prices, trade, general business conditions, etc., are fully discussed and illustrated. All of the chapters have been carefully revised, and an Appendix added. The latter includes a table of *Powers, Roots, and Reciprocals,* and a table of *Four-Place Common Logarithms.* Indeed, in its present form the book may be called new.

For suggestions and assistance in the revision, I am indebted to the students of Northwestern University who, during the past eight years, have constituted a laboratory in which the pedagogical problems of instruction in statistics and statistical methods have been observed; to instructors of statistics in other universities and to practicing statisticians with whom I have discussed the subject matter, to Professor A. L. Bowley of the London School of Economics and Political Science, who was kind enough to read in manuscript the first eight chapters, and to discuss with me, personally and at length, the different phases of statistical methods; to Professor G. Udny Yule, Cambridge University, England, Professor D. Caradog Jones, University of Liverpool, and a number of others from whom I received valuable suggestions while studying in English universities the contents of courses in Statistics and the methods of instruction; to E. J. Moulton, Professor of Mathematics, Northwestern University, who read the revision in manuscript; and to Miss Blanche L. Altman, Lecturer in Statistics, Northwestern University, and Miss Gretchen Seibert, my Secretary, both of whom assisted in the laborious task of preparing the matter for publication and in seeing it through the press.

HORACE SECRIST.

June 1, 1925.

# PREFACE TO FIRST EDITION

THE following chapters are an attempt to work out an introductory, but at the same time a comprehensive, text on statistical methods for the use of college students and students in colleges of business administration. They are also intended to supply the need for a fundamental treatment of the methods of statistical investigation and interpretation. Statistical methods are regarded as means rather than as ends, as constituting simply one phase of general methodology, and as including not only methods of analyzing but also of collecting and assembling statistical data. The methods discussed are of general application although the illustrations, for the most part, are drawn from economic and business fields.

The order of treatment is the same as that followed in the planning and analysis of a statistical problem, and it is hoped that statisticians, business executives, and students of statistical methods generally will find the volume not only a compendium of statistical procedure but also a guide in the process of logical statistical analysis. Emphasis is given to the necessity of a clear formulation of the problem in mind, to the meaning, collecting, and assembling of data, and to the necessity of a rigid interpretation and use of units of measurements. All of these steps are held to be preliminary but indispensable to the formulation of a statistical judgment, and to the employment of the refinements of mathematical analysis which alone are too generally associated with "statistical methods."

The treatment is non-mathematical for several reasons, chief of which are, that the mathematical phases of the subject are treated in other places, and that there seems to be an urgent need for a fundamental discussion of the non-mathe-

matical, but not less vital, processes in statistical investigation and analysis. Experience in teaching statistics both to college students and business men, as well as in conducting statistical investigations, has demonstrated the need for such a treatment. It has been the aim at every stage of the discussion to develop the "why" of statistics, and concretely to relate methods to the problems of public and private economics.

The bibliographical aids at the close of the several chapters are not meant to be inclusive, but are chosen because of their value to students and others as collateral reading. A discussion of certain of them along with the text treatment, and in the light of the laboratory problems assigned, has proved helpful in the author's classes.

I am indebted to Professor Willard E. Hotchkiss, formerly Dean of the Northwestern University School of Commerce, and to Professor John F. Hayford, Dean of the Northwestern University College of Engineering, for reading parts of the manuscript and for offering many helpful suggestions for its improvement. Most of all I am indebted to my wife, who has materially lightened the burden of proofreading, and who, at all stages in the preparation of the volume, has been a constant source of encouragement.

HORACE SECRIST.

November, 1917.

# CONTENTS

## CHAPTER I

## THE MEANING AND APPLICATION OF STATISTICS AND STATISTICAL METHODS

PAGE

I. Introduction . . . . . . . . . . . . . . . . 1
II. The Meaning of Statistics and Statistical Methods . 9
III. The Use and Application of Statistics and Statistical Methods . . . . . . . . . . . . . . . 13
1. Application to Individual Business Units . . . . . 15
2. Application to Groups of Business Units . . . . . 15
3. Application to Matters of General Business Growth, Decline and Change . . . . . . . . . . . 16
4. Application to Questions of Social Economy . . . . 16
5. Application to Affairs Pertaining to Governmental Discrimination and Policy . . . . . . . . . . . 16
6. Application to Questions of Economic Theory . . . 17
**References** . . . . . . . . . . . . . . . . 20

## CHAPTER II

## TYPES OF SECONDARY STATISTICAL DATA AND TESTS FOR THEIR USE

I. Introduction . . . . . . . . . . . . . . . . 22
II. Primary and Secondary Data Defined and Contrasted . 24
III. Sources of Secondary Statistical Data . . . . . . 26
IV. Tests to Be Applied to Secondary Statistical Data Before They Are Used . . . . . . . . . . . . 30
1. The Organization Supplying Secondary Data . . . . 30
2. The Purpose for Which Secondary Data Are Issued and the Consumers to Whom They Are Addressed . . 31
3. The Nature of the Secondary Data Themselves . . . 31
4. In What Types of Units Are the Data Expressed? Are They the Same at Different Times, at Different Places, and for All Cases at the Same Time or Place? 35

PAGE
5. Are the Data Accurate? . . . . . . . . . . . 38
6. Do the Data Refer to Homogeneous Conditions? . . 42
7. Are the Data Germane to the Problem Being Studied? . 44
**References** . . . . . . . . . . . . . . . . . 46

## CHAPTER III

## COLLECTING AND EDITING PRIMARY STATISTICAL DATA

I. Introduction . . . . . . . . . . . . . . . . . 47
II. Preliminary Conditions to the Collection of Primary Data . . . . . . . . . . . . . . . . 47
1. What Is the Precise Problem Upon Which Statistics Are Required? . . . . . . . . . . . . . 48
2. Does the Problem, as Formulated, Lend Itself to Statistical Treatment? . . . . . . . . . . . 48
3. What Types of Data Are Necessary for Its Analysis or Solution? . . . . . . . . . . . . . . . 49
4. Are They Likely to Be Available in Suitable Form? . 50
5. Are They Likely to Be Adequate for the Purpose in Mind? . . . . . . . . . . . . . . . . 50
6. Will They Have the Required Degree of Accuracy, Consistency, and Comparability? . . . . . . . . 51
7. Can the Data Be Made Available Within the Time Limit Required: That Is, Will They Have the Required Currency? . . . . . . . . . . . . . . . . 51
8. Are There Likely to Be Any Restrictions Upon the Use of Data Which Will Compromise the Purpose Which They Are to Serve? . . . . . . . . . . . . 52
9. What Sanction Is Necessary, and What Method of Procedure, with the Sanction Available, Must Be Followed in Order to Secure the Desired Facts? . . 52
III. The Collection Process . . . . . . . . . . . . 54
1. Purpose and Plan . . . . . . . . . . . . . 54
2. The Collection Process Descriptively Considered . . 54
3. The Collection Process Functionally Considered . . . 61
(1) Who Are to Be Canvassed? . . . . . . . . . 61
(2) The Ways in Which Primary Data May Be Secured 65
a. Personal Interviews . . . . . . . . . . . 65
b. The Use of Form or of Personal Letters . . . 65
c. The Form, Use, and Editing of Questionnaires or Schedules . . . . . . . . . . . . . 66
IV. Conclusion . . . . . . . . . . . . . . . . 70
**References** . . . . . . . . . . . . . . . . . 71

## CHAPTER IV

## UNITS OF MEASUREMENT, OF ANALYSIS, AND OF PRESENTATION IN STATISTICAL STUDIES

PAGE
I. THE MEANING OF STATISTICAL UNITS OF MEASUREMENT . . 72
II. STATISTICAL UNITS OF MEASUREMENT CLASSIFIED AND DESCRIBED . . . . . . . . . . . . . . . . . 77
1. Units of Enumeration or Estimation . . . . . . . 77
2. Units of Analysis and of Interpretation . . . . . . 80
(1) Ratios or Coefficients Relating to Time . . . . 80
(2) Ratios or Coefficients Relating to Space . . . . 81
(3) Ratios or Coefficients Relating to Condition . . . 82
III. STATISTICAL UNITS OF PRESENTATION . . . . . . . . 85
IV. DIAGRAMMATIC SCHEME ILLUSTRATING DIFFERENT TYPES OF STATISTICAL UNITS . . . . . . . . . . . . . 88
V. A SELECTED LIST OF UNITS OF ANALYSIS AND OF INTERPRETATION—*RATIOS* OR COEFFICIENTS . . . . . . . . 89
VI. RULES FOR THE USE OF STATISTICAL UNITS OF MEASUREMENT AND OF PRESENTATION . . . . . . . . . . . . 90
1. Units of Measurement . . . . . . . . . . . . 90
2. Units of Presentation . . . . . . . . . . . . 90
**References** . . . . . . . . . . . . . . . . . 91

## *CHAPTER V*

## PURPOSES OF A STATISTICAL STUDY OF WAGES, UNITS OF MEASUREMENT, SOURCES OF DATA, SCHEDULE FORMS—ILLUSTRATIONS OF METHODS

I. THE PROBLEM IN THE STUDY OF WAGES STATED . . . . . 92
1. Introduction . . . . . . . . . . . . . . . . 92
2. Characteristic Confusions in the Use of the Term "Wages" . . . . . . . . . . . . . . . 94
3. Bases for a Definition of Wages . . . . . . . . . 95
4. Wages Defined . . . . . . . . . . . . . . . 96
5. *Studies of Wages* and the Uses of *Terms* . . . . . 97
II. THE RELATION OF THE PROBLEM AS OUTLINED TO STATISTICS OF WAGES . . . . . . . . . . . . . . . . 100
1. Sources for Primary Data in Wage Studies . . . . . 100
(1) Primary Data Directly Applicable to Studies of Wages . . . . . . . . . . . . . . . 100
a. Data from Employes . . . . . . . . . . 101
b. Data from Employers . . . . . . . . . . 101
c. Data from Trade and Labor Unions . . . . 102
(2) Data Indirectly Applicable to Studies of Wages . 103

PAGE
2. Types of Secondary Wage Data . . . . . . . . . 104
(1) Secondary Data Directly Applicable to Studies of Wages . . . . . . . . . . . . . . . 104
a. Data from Employes . . . . . . . . . . 104
b. Data from Employers . . . . . . . . . . 106
(a) Material Directly Related to Wages . . . 106
(b) Material Indirectly Related to Wages . . 108
c. Data from Trade and Labor Unions . . . . 109
III. A STUDY OF WAGES: DECLARATION OF PURPOSE, DEFINITION OF UNITS, SCHEDULE FORMS . . . . . . . . . . 117
1. Declaration of Purposes . . . . . . . . . . . 117
2. Schedule and Explanation . . . . . . . . . . . 119
**References** . . . . . . . . . . . . . . . . 123

## CHAPTER VI

## CLASSIFICATION—TABULAR PRESENTATION

I. INTRODUCTION . . . . . . . . . . . . . . . . . 124
II. THE CHARACTERISTICS OF DATA TO BE TABULATED . . . . 124
1. What Are the Characteristics of Any Body of Data? . 125
2. In What Way or Ways Are the Characteristics Related to Each Other? . . . . . . . . . . . . 125
3. Can Data Be Expressed in Series with Respect to Time, Space, or Condition? . . . . . . . . . . 126
4. Are Some Characteristics Cumulative While Others Are Not? . . . . . . . . . . . . . . . 126
III. THE NATURE OF CLASSIFICATION . . . . . . . . . . 126
IV. THE MEANING OF TABULATION . . . . . . . . . . 128
V. THE ADVANTAGES OF TABULAR OVER NON-TABULAR ARRANGEMENT . . . . . . . . . . . . . . . . . 131
1. The Order of Arrangement or the Plan of Presentation . 132
(1) Arrangement According to the Size or Frequency of the Items . . . . . . . . . . . . . 132
(2) Arrangement According to Time . . . . . . 133
(3) Arrangement According to Space . . . . . . 133
(4) Arrangement According to a Variable Condition . 134
(5) Arrangement According to Alphabet . . . . . 135
2. Tabulated Data Can Be More Easily Remembered than Those Which Are Not Tabulated . . . . . . . 136
3. Visualization of Group Relations Is Facilitated . . . 137
4. A Tabular Arrangement Makes It Easy to Compare Data of Like Character . . . . . . . . . . 137
5. A Tabular Arrangement Facilitates the Summation of Items and Detection of Errors and Omissions . . 137

PAGE

6. A Tabular Arrangement Makes It Unnecessary to Repeat Explanatory Phrases and Headings . . . . . 138
VI. Types of Statistical Tables . . . . . . . . . . . 138
VII. The Tabulation Form . . . . . . . . . . . . . 139
1. Tables Classified According to Their Complexity . . 139
2. Table Structure . . . . . . . . . . . . . . 143
(1) Ruling and Spacing of Major and Minor Headings 144
(2) The Positions of Totals . . . . . . . . . 144
(3) Size of Tables and Suitability to the Printed Page . 145
(4) The Numbering of Columns and Lines . . . . 146
VIII. The Contents of Tables . . . . . . . . . . . 146
IX. Titles for Statistical Tables . . . . . . . . . 148
X. The Mechanics of Tabulation . . . . . . . . . 151
XI. Types of Statistical Series and Corresponding Tables . 157
XII. Conclusion . . . . . . . . . . . . . . . . 169
**References** . . . . . . . . . . . . . . . 170

CHAPTER VII

DIAGRAMMATIC PRESENTATION

I. Introduction . . . . . . . . . . . . . . . 171
II. Diagrams for Illustrating Frequency or Magnitude Alone . . . . . . . . . . . . . . . . 175
1. Alternative Types—Good and Bad Features of Each . 175
2. Examples of Statistical Diagrams in Current Use . . 186
3. General Rules to Be Observed in the Use of Statistical Diagrams . . . . . . . . . . . . . . 198
III. Diagrams for Illustrating Frequency or Magnitude in Relation to Spatial Distribution . . . . . . . 201
1. The Psychological Bases for the Use of Statistical Maps 201
2. Types of Statistical Maps . . . . . . . . . . 203
(1) Colored Maps . . . . . . . . . . . . 203
(2) Cross-hatched Maps . . . . . . . . . . . 204
(3) Dot Maps . . . . . . . . . . . . . . 206
IV. Suggestions to Be Followed in the Use of Statistical Diagrams . . . . . . . . . . . . . . . 212
**References** . . . . . . . . . . . . . . . 213

CHAPTER VIII

GRAPHIC PRESENTATION

I. Introduction . . . . . . . . . . . . . . . 214
II. Diagrammatic and Graphic Presentation Contrasted . 218
III. Graphic Presentation of Frequency Series . . . . . 221

PAGE
1. Plotting Simple Frequency Series . . . . . . . . 221
(1) Plotting Simple Frequency Distributions of Discrete Series . . . . . . . . . . . . . 222
(2) Plotting Simple Frequency Distributions Describing Continuous Series . . . . . . . . . . . 227
2. Plotting Cumulative Frequency Series . . . . . . 232
(1) "Graphic" Representation of Discrete Frequency Series Cumulated . . . . . . . . . . . 233
(2) Graphic Representation of Continuous Frequency Series Cumulated . . . . . . . . . . . 237
IV. Graphic Presentation of Historical or Time Series . . 242
1. Plotting Simple Historical Series . . . . . . . 243
(1) Choice and Adjustment of Scales . . . . . . 243
a. Natural Scale or "Difference" Charts . . . . 244
b. Ratio Scales and "Ratio" Charts . . . . . 248
(2) Types of Lines Connecting Successive Ordinates . 255
(3) Purposes and Methods of Smoothing Historical or Time Series . . . . . . . . . . . . 256
2. Plotting Cumulative Historical or Time Series . . . 258
V. Conclusion . . . . . . . . . . . . . . . 259
**References** . . . . . . . . . . . . . . . 259

## CHAPTER IX

## AVERAGES AS TYPES

I. Introduction . . . . . . . . . . . . . . . 261
II. Common Averages Defined . . . . . . . . . . . 263
III. The Arithmetic Mean or Average . . . . . . . . . 264
1. What It Is . . . . . . . . . . . . . . . 264
2. How the Arithmetic Mean Is Computed . . . . . 267
3. Some "Do's and Don'ts" in the Use of Averages . . . 278
4. Summary . . . . . . . . . . . . . . . 282
IV. The Median . . . . . . . . . . . . . . . 282
1. What the Median Is . . . . . . . . . . . . 282
2. How the Median Is Determined . . . . . . . . 283
3. Summary . . . . . . . . . . . . . . . 294
V. The Mode . . . . . . . . . . . . . . . 294
1. What the Mode Is . . . . . . . . . . . . 294
2. How the Mode Is Located . . . . . . . . . . 297
(1) The Location of the Mode in Historical or Time Series . . . . . . . . . . . . . 297
(2) The Location of the Mode in Space Series . . . 300
(3) The Location of the Mode in Frequency Series . . 300
3. Summary . . . . . . . . . . . . . . . 307

PAGE
VI. THE GEOMETRIC MEAN . . . . . . . . . . . . . 307
VII. THE PROPERTIES OF THE ARITHMETIC MEAN, THE MEDIAN, THE MODE, AND THE GEOMETRIC MEAN COMPARED AND CONTRASTED . . . . . . . . . . . . . . . 310
VIII. THE AVERAGE TO USE—SOME TYPICAL CASES WHERE CHOICE IS IMPORTANT . . . . . . . . . . . . . . 311
IX. SUMMARY AND CONCLUSION . . . . . . . . . . . . 320
**References** . . . . . . . . . . . . . . . . 322

## CHAPTER X

## DISPERSION

I. INTRODUCTION . . . . . . . . . . . . . . . 324
II. THE MEANING OF DISPERSION . . . . . . . . . . . 325
III. MEASURES AND COEFFICIENTS OF DISPERSION . . . . . 326
1. The Method of Limits . . . . . . . . . . . 326
(1) The Range . . . . . . . . . . . . . 326
(2) The Decile Method (Graphic) for Time Series . . 331
2. The Method of Averaging Differences from a Type . 336
(1) The Average Deviation . . . . . . . . . 337
a. The Average Deviation in Historical Series . . 339
b. The Average Deviation in Frequency Series . . 342
(2) The Standard Deviation . . . . . . . . . 349
a. The Standard Deviation in Historical or Time Series . . . . . . . . . . . . . 352
b. The Standard Deviation in Frequency Series . . 354
(3) The Quartile Measure . . . . . . . . . . 356
IV. SUMMARY . . . . . . . . . . . . . . . . 358
**References** . . . . . . . . . . . . . . . . 358

## CHAPTER XI

## THE THEORY OF PROBABILITY AND SOME PROPERTIES OF THE NORMAL LAW OF ERROR DISTRIBUTION

I. OUTLINE OF THE THEORY OF PROBABILITY . . . . . . 360
II. PROPERTIES OF THE NORMAL LAW OF ERROR DISTRIBUTION . 367
III. THE MEANING OF THE PROBABLE ERROR CONCEPT . . . . 370
IV. SAMPLE MEASUREMENTS AND THE USES OF PROBABLE ERROR . 372
V. SUMMARY . . . . . . . . . . . . . . . . 374
**References** . . . . . . . . . . . . . . . . 374

## CHAPTER XII

## SKEWNESS OR ASYMMETRY

PAGE

I. Introduction . . . . . 376
II. Dispersion and Skewness Contrasted . . . . . 377
III. Types of Skewed Distributions . . . . . 378
IV. Measures and Coefficients of Skewness . . . . . 381
V. Methods of Summarizing Frequency Series . . . . . 386
VI. Conclusion . . . . . 391
**References** . . . . . 391

## CHAPTER XIII

## THE THEORY AND MEASUREMENT OF CORRELATION

I. Introduction . . . . . 393
II. Comparison, Causation, and Correlation . . . . . 394
III. The Meaning of Correlation . . . . . 398
1. Definition and Explanation . . . . . 398
2. Illustrations of Correlation by Throws of Dice . . . 400
IV. The Measurement of Correlation . . . . . 406
1. The "Sum Product" Method . . . . . 406
(1) The Assumptions Upon Which the Pearsonian Coefficient of Correlation Is Based . . . . . 406
(2) The Pearsonian Coëfficient of Correlation Formula 410
(3) The Calculation of the Pearsonian Coefficient of Correlation . . . . . 413
a. In Ungrouped Series . . . . . 413
b. In Grouped Series . . . . . 417
(4) Regression Lines and Coefficients of Regression . 425
(5) The Probable Error of the Coefficient of Correlation 428
2. The Concurrent Deviation Method . . . . . 430
3. Graphic Methods of Showing Association Between Different Variables . . . . . 432
V. Conclusion . . . . . 433
**References** . . . . . 436

## CHAPTER XIV

## THE TREATMENT AND CORRELATION OF TIME SERIES

I. Introduction . . . . . 438
II. The Nature of Changes in Time Series . . . . . 438

PAGE
III. Methods of Measuring and Isolating Time Changes . . 441
1. Methods of Measuring Long-time or Secular Trend . . 442
(1) The Free-Hand Method . . . . . . . . . 443
(2) The Method of Averaging . . . . . . . . 444
(3) The Least-Square Method . . . . . . . . 444
2. Methods of Measuring Normal Seasonal Change . . . 448
(1) Monthly Means or Averages . . . . . . . . 449
(2) The Method of Moving Medians . . . . . . 449
(3) The Median-Link-Relative Method . . . . . 450
3. Cyclical Fluctuations . . . . . . . . . . . 455
IV. The Correlation of Time Series . . . . . . . . 457
V. The Probable Error of the Correlation Coefficient of Time Series . . . . . . . . . . . . . . 464
VI. Conclusion . . . . . . . . . . . . . . . 465
References . . . . . . . . . . . . . . . 466

CHAPTER XV

THE PRINCIPLES OF INDEX NUMBER MAKING AND USING

I. Introduction . . . . . . . . . . . . . . . 468
II. Index Numbers Defined and the Methods of Computing Them Illustrated . . . . . . . . . . . . 469
1. The Average of Relatives (Ratios) Method . . . . 470
(1) "Simple" Average of Relatives (Ratios) . . . . 470
a. Fixed Base . . . . . . . . . . . . 470
b. Chain Base . . . . . . . . . . . . 472
(2) Weighted Average of Relatives (Ratios) . . . . 473
2. Ratios of Averages . . . . . . . . . . . . 476
3. Ratios of Weighted Aggregates . . . . . . . . 477
4. Summary of Results by Different Methods . . . . 479
III. The Uses of Index Numbers . . . . . . . . . . 480
IV. Principles of Index Number Making . . . . . . . 481
1. The Attributes of Index Numbers and the Steps in Their Construction . . . . . . . . . . . . . 482
2. Data from Which Price Index Numbers Are Made . . 484
3. Dispersion of Price Fluctuations . . . . . . . . 489
V. The Methods of Constructing Index Numbers . . . 496
1. Averages of Relatives (Ratios) . . . . . . . . 496
(1) Fixed vs. Shifting Base . . . . . . . . . . 496
a. Arithmetic Means of Relatives—Fixed Base . . 497
b. Medians of Relatives—Fixed Base . . . . . 497
c. Geometric Means of Relatives—Fixed Base . . 498

PAGE
(2) Chain Relatives . . . . . . . . . . . . . . 499
(3) Base Shifting and the Use of Averages of Relatives 500
a. When Arithmetic Averages of Relatives Are Used 500
b. When Medians of Relatives Are Used . . . . 501
c. When Geometric Means of Relatives Are Used . 501
2. Ratios of Average Prices . . . . . . . . . . . 502
(1) Merits of the Method . . . . . . . . . . . 502
(2) Methods of Base Shifting Illustrated . . . . . 503
3. Weighted Aggregates of Actual Prices and Base Shifting 504
(1) Method of Computation and Relative Merits . . 504
(2) Methods of Base Shifting Illustrated . . . . . 504
VI. Weighting . . . . . . . . . . . . . . . . . . 505
1. Meaning and Methods of Weighting . . . . . . . 505
2. Weighting in Professor Fisher's "Ideal" Formula . . . 509
VII. Suggestions to Users of Price Index Numbers . . . . 511
VIII. Conclusion . . . . . . . . . . . . . . . . 513
References . . . . . . . . . . . . . . . . . 513

CHAPTER XVI

PRICE, QUANTITY, AND GENERAL BUSINESS INDEXES DESCRIBED AND COMPARED

I. Introduction . . . . . . . . . . . . . . . . 515
II. Index Numbers of Prices . . . . . . . . . . . . 515
1. Price Index Numbers Issued by the United States Government . . . . . . . . . . . . . . . . 516
(1) Index Numbers of Wholesale Prices . . . . . . 516
a. The United States Bureau of Labor Statistics' Wholesale Price Index Number . . . . . 516
b. The Federal Reserve Board's Wholesale Price Index Number . . . . . . . . . . . . 518
c. The United States Department of Agriculture's Wholesale Price Index Number of Farm Prices of Crops and of Livestock . . . . . . . . 519
(2) Index Numbers of Retail Prices . . . . . . . 520
a. The United States Bureau of Labor Statistics' Index Number of Food Prices . . . . . . 520
b. The United States Bureau of Labor Statistics' Index Number of Cost of Living . . . . . 521
2. Wholesale Price Index Numbers Issued by Private Organizations . . . . . . . . . . . . . . 523
(1) Bradstreet's Index Number . . . . . . . . . 523
(2) Dun's Index Number . . . . . . . . . . . 525
(3) The New York Annalist's Index Number . . . . 527

PAGE
(4) Professor Fisher's Index Number . . . . . . 528
(5) The Commodity Price Index of Business Cycles of the Harvard Committee on Economic Research 529
III. INDEX NUMBERS OF PRODUCTION . . . . . . . . . . 530
1. The Index of Physical Production of the Harvard Committee on Economic Research . . . . . . . 531
(1) Index of Agricultural Production . . . . . . 531
(2) Index of Mining . . . . . . . . . . . . . 532
(3) Index of Manufacture . . . . . . . . . . . 533
(4) Combined Index of Agriculture, Mining, and Manufacture . . . . . . . . . . . . . . . . 534
2. Other Indexes of Physical Production . . . . . . 535
(1) The Federal Reserve Board . . . . . . . . . 535
(2) The Department of Commerce . . . . . . . . 535
IV. INDEXES OF VOLUME OF TRADE . . . . . . . . . . . 536
1. "Persons'" Index of the Harvard Committee on Economic Research . . . . . . . . . . . . . 536
2. "Snyder's" New Index of the Volume of Trade . . . 536
3. Other Trade Indexes . . . . . . . . . . . . . 537
(1) The Federal Reserve Board . . . . . . . . . 537
(2) The United States Department of Commerce . . 537
V. INDEXES OF GENERAL BUSINESS CONDITIONS . . . . . . 537
1. The Harvard Index of General Business Conditions . . 538
2. The Brookmire Forecasting Composite Line . . . . 541
3. Other Barometric and Forecasting Indexes . . . . 543
VI. OTHER INDEXES OF BUSINESS AND ECONOMIC PHENOMENA . 543
1. Money and Prices . . . . . . . . . . . . . . 544
2. Employment and Unemployment . . . . . . . . 544
3. Index of Foreign Exchange Rates . . . . . . . . 545
4. Indexes of Distribution . . . . . . . . . . . 545
5. Indexes of Security Prices . . . . . . . . . . . 545
6. Indexes of Earnings and Wage-Rates . . . . . . . 545
VII. CONCLUSION . . . . . . . . . . . . . . . . . 546

APPENDIX . . . . . . . . . . . . . . . . . . . . 548
Table 1—Powers, Roots, Reciprocals . . . . . . . . . 548
Table 2—Common Logarithms—Four Places . . . . . . . 566
INDEXES . . . . . . . . . . . . . . . . . . . . . 569
Subject . . . . . . . . . . . . . . . . . . . . 569
Personal . . . . . . . . . . . . . . . . . . . . 583

# LIST OF FIGURES

FIGURE PAGE

1. Diagrammatic Scheme Illustrating Different Types of Statistical Units . . . . . . . . . . . . . . . . 88
2. Hand Tabulation Card . . . . . . . . . . . . . 153
3. Machine Tabulation Card . . . . . . . . . . . . 154

4–21. Different Types of Lines, Bars, Squares, Cubes, and Circles Diagrammatically Illustrating Data . . . . . . . . 176–186

22. Diagram Showing Discrete Time Series . . . . . . . . 187
23. Diagram Showing a Discrete Space Series . . . . . . . . 188
24. Diagram Showing a Discrete Space Series . . . . . . . 188
25. Value of Petroleum and Natural Gas, by States, 1909 (Illustrations of Lines, Surfaces, and Volumes) . . . . . . . 189
26. Public School Property in 1903 and 1914 (Solids Drawn Out of Scale) . . . . . . . . . . . . . . . . . . . . 190
27. Payments, Account Bonded Debt and Interest, on County Bonds (Solids Drawn Out of Scale) . . . . . . . . . 190
28. Diagram Showing a Discrete Condition Series . . . . . . 191
29. Diagram Showing Component Parts—Discrete Time Series . 192
30. Diagram Showing Component Parts—Discrete Time Series . 192
31. Pie Diagram Showing Component Parts . . . . . . . . 193
32. Pie Diagram Showing Component Parts—Percentages of Expenditures for Major Items of Family Budget . . . . 194
33. Pie Diagrams Showing Component Parts by Years . . . 194
34. Production of Petroleum, by Fields, 1909 (Sectors of Circles and Lines) . . . . . . . . . . . . . . . . . . . . 195
35. Two-Dimensional Bar Diagram Showing Discrete Condition Series . . . . . . . . . . . . . . . . . . . . . 196
36. Color or Race, Nativity, and Parentage, by Divisions of the United States, 1910 . . . . . . . . . . . . . . . . 197
37. Two-Dimensional Diagram Showing Components by Use of Surfaces Within Surfaces . . . . . . . . . . . . . 198
38. Proportion of Males 10 to 13 Years of Age Engaged in Gainful Occupations, by States, 1910 (Cross-hatched Map) . . 205
39. Primary Markets for Wisconsin Cheese (American), 1911 . 207
40. Pig Iron Production, by States, 1909 . . . . . . . . . 209
41. Number of Swine on Farms and Ranges, April 15, 1910 (1 Dot = 2500) . . . . . . . . . . . . . . . . . . . . 211
42. Bar Diagrams Showing the Number of Employes in Factory "X" Classified by Age . . . . . . . . . . . . . 216

FIGURE PAGE

43. Bar Diagram Showing Hourly Temperature Readings at Chicago, September 3, 1924 . . . . . . . . . . . . . 218
44. A System of Co-ordinates . . . . . . . . . . . . 219
45. Smoothed Frequency Distribution of Lengths of Ears of Corn (Frequency Distribution, Continuous Series) . . . . . 231
46. Bar Diagram Showing a Discrete Frequency Series Cumulated on a "Less Than" Basis . . . . . . . . . . . 235
47. Bar Diagram Showing a Discrete Frequency Series Cumulated on a "More Than" Basis . . . . . . . . . . . 236
48. Cumulative Graphs—Ogives—Constructed on "More Than" and "Less Than" Bases, Showing by Towns the Classified Prices of Oil . . . . . . . . . . . . . . . . 239
49. Cumulative Graph of a Continuous Frequency Series Showing Length of Time Taken to Thread a Standard Bolt (Basis of Cumulation—"Less Than") . . . . . . . . . . 241
50. Capital and Clearings of New York Clearing House Banks, 1902-1915 (Method of Scale Conversion) . . . . . . . 246
51. Capital and Clearings of New York Clearing House Banks, 1902-1915 (Method of Scale Conversion) . . . . . . . 247
52. A Natural or Difference Scale Contrasted with a Percentage or Ratio Scale . . . . . . . . . . . . . . . . 249
53. Illustration of How Different Scales May Be Placed on a Ratio Background . . . . . . . . . . . . . . 250
54–55. Difference and Ratio Charts Showing the Changes in Funds "A" and "B" . . . . . . . . . . . . . . 251
56–57. Difference and Ratio Charts Showing the Changes in Volume of Sales of Three Products . . . . . . . . . 252
58. Domestic Orders for Freight Cars and Locomotives, Plotted on a Ratio Chart . . . . . . . . . . . . . . 253
59. Rate of Turnover of Bank Deposits, Plotted on a Ratio Chart 254
60. Exports and Domestic Consumption of Cotton, Plotted on a Ratio Chart . . . . . . . . . . . . . . . . 254
61. Diagrams Illustrating the Nature of the Arithmetic Mean When Items Are Differently Weighted . . . . . . . . 268
62. Cumulative Graphs—Historigrams—Constructed on "Up to and Including" and "After and Including" Bases, Showing by Years, Importations of Raw Cotton into the United States . . . . . . . . . . . . . . . . . . . 292
63. Histograms Showing the Distributions of Ratios of Assessed Values of Buildings to the Assessed Values of Lands upon Which They Stand, New York City, 1914 . . . . . . 306
64. Curves Showing, by the Range and the Decile Methods, the Dispersion of the Fluctuations in Relative Wholesale Prices of 145 Commodities, 1890-1910 . . . . . . . . . . 333
65. Conspectus of Yearly Changes in Prices, 1891-1918 . . . . 335
66. Types of Frequency Distributions . . . . . . . . . . 343

FIGURE PAGE

67. Graphical Representation of the Theoretical Distribution Secured by Tossing Ten Coins . . . . . . . . . . 368
68. The Area of the Normal Curve, Inside (blank), and Outside (shaded), the Limits Set by One Times the Probable Error 370
69. The Area of the Normal Curve, Inside (blank), and Outside (shaded), the Limits Set by Twice the Probable Error . . 371
70. The Area of the Normal Curve, Inside (blank), and Outside (shaded), the Limits Set by Three Times the Probable Error . . . . . . . . . . . . . . . . . . 372
71. The Form of the Ideal Symmetrical Frequency Distribution 378
72. The Forms of Ideal Moderately Asymmetrical or Skewed Distributions . . . . . . . . . . . . . . 379
73. U-Shaped Distribution Curve of Deaths per 1,000 Population at Specified Age Periods, United States Registration States, 1920 . . . . . . . . . . . . . . . . . . 380
74. The Form of the Ideal J-Shaped Frequency Distribution Curve . . . . . . . . . . . . . . . . . 381
75. Curves Showing, by Years, Classified Wage-Rates of Female Menders in Woolen and Worsted Establishments, 1907-1908 389
76. Curves Showing, by Years, Classified Wage-Rates of Female Menders in Woolen and Worsted Establishments, 1909-1910 390
77. Graphic Figures Illustrating Correlation by Means of 500 Pairs of Throws of Dice . . . . . . . . . . . 405
78. Regression Lines of Rent per Unit of Floor Space on Rent per Unit of Sales, and Rent per Unit of Sales on Rent per Unit of Floor Space for 150 Retail Clothing Stores . . . 426
79. Amounts of Inventory per $100 of Total Net Sales for Stores Classified by Size, 1919, 1918, and 1914, Combined . . . 433
80. Annual Rates of Stock Turnover for Stores Classified by Size, 1919 . . . . . . . . . . . . . . . 434
81. Amounts of Wages and Salaries per $100 of Total Expense for Stores Classified by Size, 1919, 1918, and 1914, Combined . 435
82. Curves Showing Long-time or Secular Changes . . . . . 439
83. Chart Showing the Actual Production of Pig Iron in the United States, 1903 to 1916, and a Line Showing the Long-time Trend . . . . . . . . . . . . . . . 442
84. Pig Iron Production, 1903-1916—Figures Corrected for Long-time Trend (Percentages) . . . . . . . . . . . 448
85. Pig Iron Production—Percentages—Corrected for both Secular Trend and Seasonal Variation . . . . . . . . . 453
86. Distribution of the Price Variations of 241 Commodities in 1913 (Percentages of Rise or Fall in Prices) . . . . . 490
87. Distribution of 5578 Price Variations (Percentages of Rise or Fall from Prices of Preceding Year) . . . . . . . 494
88. The Harvard Index of General Business Conditions, 1903-1914 540
89. The Harvard Index Chart of General Business Conditions, 1919-1924 . . . . . . . . . . . . . . . 542

# LIST OF TABLES

TABLE PAGE

1. Table Showing by Sex the Nature of Changes in an Employed Force in Factory "A," 1923 and 1924 . . . . . . . . 131
2. Table Showing the Names of Industries and Numerical Ranking by Value of Product (United States Census of Manufactures, 1909) . . . . . . . . . . . . . . 132
3. Number of Employes of Railroads in Service June 30, 1913 . 136
4. Railway Freight Cars, Number in Service, 1913 . . . . . 136
5. Developed Water Power Resources, Horse-Power, 1900, by Drainage Basins . . . . . . . . . . . . . . 136
6. Number of Deaths in the United States by Causes, 1913 . . 136
7. Table Showing by Years the Number of Real Estate Mortgages in Wisconsin . . . . . . . . . . . . . . 140
8. Table Showing by Years the Number of Real Estate Taxable and Non-Taxable Mortgages in Wisconsin . . . . . . 141
9. Table Showing by Years the Number and Amount of Real Estate Taxable and Non-Taxable Mortgages in Wisconsin 141
10. Table Showing by Years and by Districts of the State the Number and Amount of Taxable and Non-Taxable Real Estate Mortgages in Wisconsin . . . . . . . . . 142
11. The Causes of Accidents Resulting in Infection . . . . . 150
12. Jointer Accidents Reported, by Nature of Disability . . . 150
13. Accidents Caused by Falls of Workmen—by Cause and Disability . . . . . . . . . . . . . . . . . 151
14. Table Showing Union Scales of Wages for Plumbers on October 1, 1913, by Municipalities . . . . . . . . . . . 158
15. Frequency Table Showing Classified Weekly Wages for Employes in All Manufacturing Industries in Massachusetts, 1912 . . . . . . . . . . . . . . . . . 160
16. Frequency Table Showing the Number of Deaths from All Causes . . . . . . . . . . . . . . . . . 161
17. Frequency Tables Showing the Number of Real Estate Mortgages in Wisconsin, 1904, Classified by Rates of Interest . 164
18. Frequency Table Showing Distribution of the Lengths of Lobsters . . . . . . . . . . . . . . . . 165
19. Frequency Table Showing the Number of the Operatives in Woolen and Worsted Mills in the United States, by Sex and by Hourly Rates of Wages . . . . . . . . . . . 168
20. Table Showing the Percentage Relation of the Assessment of Personal Property to Total Assessment . . . . . . . 169

TABLE PAGE

21. Stocks of Merchandise Illustrating Different Types of Statistical Series (Time Series) (Space Series) (Condition Series) . . . 175
22. Number of Employes in Factory "X," Classified by Age . . 215
23. Temperature Measurements at Hourly Intervals, Chicago, September 3, 1924 . . . 217
24. Proposed Freight Rates per 100 Pounds Between St. Paul, Minneapolis, and Sioux City, Iowa, Ending in Different Integers . . . 223
25. Table Showing the Number of Females and Minors Employed in 24 Mercantile Establishments in September, 1913, Receiving Classified Wage-Rates . . . 226
26. Table Showing the Number of Ears of Corn Classified by Lengths . . . 228
27. Cumulations of Weekly Wage-Rates on a "Less Than" Basis 234
28. Cumulations of Weekly Wage-Rates on a "More Than" Basis 234
29. Table Showing the Distribution of Towns According to Prices Paid for Oil, Freight Deducted (1830 Quotations), December, 1904, for the United States . . . 238–239
30. Lengths of Time Taken to "Thread" a Standard Bolt . . . 240
31. Table Showing Wage-Rates as Bases for the Computation of a Simple Arithmetic Mean Rate . . . 267
32. Table Showing Wage-Rates with Number of Persons Receiving Them as a Basis for Computing an Arithmetic Mean Rate . . . 269
33. Table Showing Wage-Rates with Number of Persons Receiving Them as a Basis for Computing Arithmetic Mean Rates . 270
34. Table Giving Data for Computing the Arithmetic Mean by the "Short-Cut" Method . . . 272
35. Table Giving Data for Computing the Arithmetic Mean by the "Short-Cut" Method . . . 273
36. Table Giving Data for Computing an Arithmetic Mean from Frequency Groups . . . 274
37. Table Giving Data for Computing an Arithmetic Mean by the "Short-Cut" Method for Frequency Groups from an Assumed Average . . . 274
38. Table Showing the Effect of Computing the Arithmetic Mean from the True Average for Data in Frequency Groups . . 275
39. Table Giving Data for Computing the Arithmetic Mean by the "Step-Deviation" Method for Frequency Groups from an Assumed Average . . . 276
40. Table Giving Data for Computing the Arithmetic Mean by the "Step-Deviation" Method from an Assumed Average When the Groups Are of Unequal Size . . . 277
41. Table Giving Data for Computing the Median . . . 283
42. Table Giving Data Showing the Effect of Changes of Distribution on the Median and the Arithmetic Mean . . . 285

TABLE PAGE

43. Table Giving Frequency Data for the Computation of the Median . . . 287
44. Table Showing by Years Singly and Cumulatively the Quantity of Raw Cotton Imported into the United States, 1895 to 1913, Inclusive . . . 291
45. Table Showing Data of Importations of Raw Cotton Arranged so as to Determine the Median Amount Imported 293
46. Data Showing Importations of Raw Cotton into the United States, Arranged so as to Determine the Modal Amount . 298
47. Number of Store Periods (Monthly) in Which Ratios of Operating Expense to Sales Were Classified Amounts in Retail Meat Stores . . . 303
48. Table Showing the Steps Used in Calculating a Geometric Mean . . . 309
49. Table Showing Deaths and Death-Rates of Married and Unmarried Men in Scotland, 1863, Classified by Age Groups . 316
50. Table Illustrating the Cumulative- or Moving-Range Method of Showing Dispersion in Historical Series . . . 327
51. Table Illustrating the Cumulative- or Moving-Range Method of Showing Dispersion in Frequency Series . . . 328
52. Table Showing the Deciles of Relative Wholesale Prices in the United States, by Years—1890-1910 . . . 332
53. Table Showing the Quantity of Tin Plates Imported into the United States, 1906-1915, Inclusive, in Millions of Pounds . 340
54. Table Showing in Classified Form the Differences from the Average Importations of Tin Plates into the United States 340
55. Table Showing the Method of Computing the Average Deviation When an Assumed Average Is Used . . . 341
56. Table Showing the Method of Computing the Average Deviation in a Simple Frequency Distribution . . . 344
57. Table Showing the Method of Computing the Average Deviation from a Group-Frequency Series . . . 345
58. Table Showing the Method of Computing the Average Deviation in a Group-Frequency Series When an Assumed Average Is Used . . . 346
59. Table Showing the Method of Computing the Average Deviation in a Group-Frequency Series from an Assumed Average by the "Step-Deviation" Method . . . 348
60. Table Showing the Method of Computing the Standard Deviation for Historical Series Using the Direct Method . . 352
61. Table Showing the Method of Computing the Standard Deviation for Historical Series Using the Direct Method but an Assumed Average . . . 353
62. Table Showing the Method of Computing the Standard Deviation for Frequency Series by Using the Short-Cut Method and an Assumed Average . . . 354
63. The Theoretical Distribution Secured by Tossing Ten Coins 366

TABLE PAGE

64. Table Showing Classified Wage-Rates of Female Menders in Eighteen Identical Woolen and Worsted Manufacturing Establishments, by Years, Together with Certain Measures of Dispersion and Skewness . . . . . . . . . . 388
65. Table Showing the Distribution of Dice with Four or More Spots Uppermost in 1,000 Throws . . . . . . . . 400
66. Table Giving the Results of 500 Pairs of Throws of 12 Dice When All Those Thrown the First Time Were Thrown the Second Time . . . . . . . . . . . . . . . . . 401
67. Table Giving the Results of 500 Connected Throws of 12 Dice, in Each Second Throw of Which 3 Dice Were Left Down and Counted . . . . . . . . . . . . . 402
68. Table Giving the Results of 500 Connected Throws of 12 Dice, in Each Second Throw of Which 5 Dice Were Left Down and Counted . . . . . . . . . . . . . 403
69. Table Giving the Results of 500 Connected Throws of 12 Dice, in the Second Throws of Which 10 Dice Were Left Down and Counted . . . . . . . . . . . . . 404
70. Table Showing by States the Capacity Load Factor and the Income per Kilowatt Hour in the Generation of Electrical Energy . . . . . . . . . . . . . . . . . 414–415
71. Table Showing the Method of Calculating the Correlation Coefficient for Grouped Series, Deviations Being Taken from the True Arithmetic Mean . . . . . . . . 418–419
72. Table Showing the Method of Calculating the Correlation Coefficient for Grouped Series, the Deviations Being Taken from an Assumed Arithmetic Mean . . . . . . . . 422–423
73. Number of Identical Retail Clothing Stores Distributed According to the Amount and Type of Their Expense Deviations from the Average in Two Successive Years . . . 431
74. Monthly Production of Pig Iron in the United States . . . 441
75. Monthly Production of Pig Iron, 1903-1916 (Showing Method of Determining Monthly Increment of Trend) . . . . 446
76. Table Showing Monthly Link Relatives of Pig Iron Production, 1903 to 1916 . . . . . . . . . . . . . . 451
77. Table Showing Actual Pig Iron Production, Least-Square Ordinates of Trend, Seasonal Variation and Cycle Percentages, 1903 to 1916; and Cycle Percentages of Interest Rates on 60-90 Day Commercial Paper, New York, 1903-1916 . . . . . . . . . . . . . . . . . . . 454
78. Average Wholesale Prices of Different Types of Paper in Chicago, 1913-1921 . . . . . . . . . . . . . . 470
79. Relative Wholesale Prices of Paper in Chicago, 1913 to 1921 471
80. Table Showing Chain-Relative Index Numbers of Wholesale Prices of Paper in Chicago, 1913 to 1921 . . . . . . . 473
81. Table Giving a Weighted Average of Relatives Index Number of Wholesale Prices of Paper in Chicago, 1913 to 1921. Base Weights: Value of Paper Consumed in 1917 . . . . . 474

TABLE PAGE

82. Table Showing Weighted Medians of Relatives Index Number of Wholesale Paper Prices, Chicago, 1913 to 1921 . . . 475
83. Table Showing the Method of Computing a Weighted Median of Relatives Index Number of Wholesale Paper Prices in Chicago, 1916 . . . . . . . . . . . . . . . . 476
84. Table Showing Ratios-of-Averages Index Numbers of Wholesale Paper Prices in Chicago, 1913-1921 . . . . . . . . 477
85. Table Giving Weighted Aggregate of Actual Prices Index Numbers of Wholesale Prices of Paper in Chicago, 1913 to 1921 . . . . . . . . . . . . . . . . . . . . 478
86. Index Numbers of Wholesale Prices of Paper in Chicago, 1913-1921, Computed by Different Methods . . . . . . . . 479
87. Distribution of 5,578 Cases of Change in the Wholesale Prices of Commodities from One Year to the Next, According to the Magnitude and Direction of the Changes . . . . . 492

AN INTRODUCTION
TO STATISTICAL METHODS

No. 11 APR 1921 COIMBATORE

# AN INTRODUCTION TO STATISTICAL METHODS

## CHAPTER I

## THE MEANING AND APPLICATION OF STATISTICS AND STATISTICAL METHODS

### I. Introduction

It is coming to be the rule to use statistics and to think statistically. The larger business units not only have their own statistical departments in which they collect and interpret facts about their own affairs, but they themselves are consumers of statistics collected by others. The trade press and government documents are largely statistical in character, and this is necessarily so, since only by the use of statistics can the affairs of business and of state be intelligently conducted.

Business needs a record of its past history with respect to sales, costs, sources of materials, market facilities, etc. Its condition, thus reflected, is used to measure progress, financial standing, and economic growth. A record of business changes —*of its rise and decline* and *of the sequence of forces influencing it*—is necessary for estimating future developments. This necessity extends not only to matters affecting accounts and accounting, but also to sales, population growth, consumer-demand, transportation, sources of raw material, advertising and display, industrial accidents and liability, capital accumulation, income distribution, marketing possibilities, prices and price movements, credit and banking facilities, production, etc.

Accounting alone does not meet this need. It is concerned

primarily with recording debtor and creditor relations and financial transactions, and with balancing accounts. These are all necessary, but they are inadequate. They do not fully disclose the workings of all phases of business, nor do they cover all aspects of business with which management is concerned. Moreover, the method takes account more of individual than group transactions. It is concerned primarily with a summation of details into totals and with the distribution of accounts and financial transactions among the respective groups of which they are a part. It does not treat with aggregates as such nor with the averages which serve to characterize them. It does not deal with the "law of large numbers"—statistical regularity—but rather with the detail out of which the aggregates are made up. Its technique and method are different from that which has come to be known as statistical methodology. How different will appear more clearly as that relating to statistics is developed in what follows.

Because it has become necessary to base economic, business, and social policies upon facts; and because the collection, use, and interpretation of such facts require the knowledge of a special technique, instruction in statistical methods is necessary. It is the main purpose of this book to serve as an introduction to such methods. That this need is keenly felt is evident from the fact that universities, almost without exception, give statistics and statistical methods a place in their curricula; and that business firms, trade and industrial associations, government bureaus, and others actively compete for the services of those whose knowledge extends to this subject.

While it is coming to be appreciated that a knowledge of facts and action based upon them are necessary as a basis for business and social policies, this point of view is not universal. It is still common for business men to base their policies upon "hunches" and hearsay. The same is true in other walks of life. Statesmen, legislators, and social workers sometimes scout "statistics," and support their beliefs and programs on a less secure foundation. These arise in tradi-

tion, customary belief, and prejudice. On the whole, however, respect for both statistics and statistical methods is deepening and broadening. What is now being done is closely to observe conditions, to enumerate the frequency with which they occur, to analyze the relations between them, and to generalize in the light of such observations. This is as it should be.

The study of statistics is largely concerned with methods—methods of collecting and utilizing numerical data in order to understand economic, business, and social problems. Its aim is to reduce to a workable basis the methods of statistical analysis, to state the principles which govern such analysis, and to illustrate the ways in which the methods may be applied to the affairs of life. It is essentially practical, yet is far more than vocational. Statistical methods, wherever applicable, are much alike. The fundamentals are the same wherever used; only in minor respects do the details differ. Their general application makes statistics a suitable subject for study.

The following treatment, while primarily keeping in mind the needs and problems of the student and of the business man, is broad enough to serve as an introduction wherever statistics are used. It is assumed that the student is scientifically inclined, that he is without prejudice, and is open-minded. It is taken for granted that he wishes to understand the problems with which he deals, to acquire a knowledge and an understanding of the methods by which problems may be approached statistically, and to acquire a certain amount of technique in dealing with them. It is also assumed that business men and others desire to act rationally upon the basis of facts, and to formulate their judgments in the light of their proper interpretation.

The statistical approach to the study of the facts of life, however, does not preclude the use of other methods. Indeed, with respect to some, it has no application. Some phenomena cannot be quantitatively measured. Honesty of purpose, re-

sourcefulness, integrity, good-will—all important in industry *as well as in life generally—are not susceptible of direct sta*tistical measurement. Where it is applicable, there is often too much faith placed in statistics alone. Statistics are used as "proof," when as a matter of fact little or nothing can be "proved" by them. What can be done by them is to describe problems quantitatively, break them up into their different parts, summarize the facts about them, and prepare the way for a logical inference. The latter, however, must be made in part on other than statistical evidence.

While statistics do not supply conclusions, they do furnish in part the basis on which they may be drawn. When "statistics" are available, however, reason is frequently dispensed with. Indeed, reasoning is sometimes thought to be equivalent to citing "statistics." The two, however, are not identical. Statistics are sometimes quoted as "proof," notwithstanding the fact that they may (1) have no application to the problem being considered, (2) be incomplete, and (3) be unrepresentative and questionable in origin. Obviously, this condition obtains when ignorance holds sway, or when design prompts one to confuse his opponent by quoting what appears to be irrefutable "statistics." Moreover, not all problems can be measured in statistical terms, nor conclusions about them be reached by the use of statistical methods. Loose reasoning nd *faulty judgments, of course, are never defensible, but* here is less excuse for them when statistics are used as "proof" han when they are ignored. This follows because statistics eem to be exact—the mere fact that they are expressed as lefinite quantities makes them appear precise. Appearance n this form, however, is a guaranty neither of accuracy nor f application.

The significant thing about statistics is not so much the umerical quantities which are attached to things counted as s is the identity of the things themselves. Indeed, the same uantitative difference does not necessarily have the same gnificance. For instance, the difference between 6 and 7

is 1. The difference between 246,789 and 246,790 also is 1, but it is not necessarily the same 1. It is certainly not the same *proportional* difference. The first may be real; the second is probably fictitious. Only as quantities are they alike; in significance they may be entirely different.

The facts of business, economic, and social life which are expressed statistically are traceable to a multitude of causes. Rarely do they stand alone as isolated occurrences. They are related to other facts. They occur in sequences with respect to time, space, or condition at a given time or space.

"A given economic fact is the result of numerous complex forces, many of which are in a state of constant variation and react upon one another; and of these forces only a few can be adequately described by the method of statistics. Consequently these few are often quoted as if they were the only active causes whereas the effect attributed to them is probable only on the assumption that all other causes remain unchanged or suspended. . . . Statistics, even when compiled accurately, though often absolutely necessary for a complete solution of a problem, do not in themselves provide that solution, but are to be used in conjunction with evidences of other kinds."[1]

The important steps involved in the use of statistics are: (1) observation, (2) measurement, (3) analysis, and (4) inference. It is the multitude of processes and methods connected with each of these steps with which this book is concerned. Because they are misunderstood or ignorantly carried out, statistics are often in disrepute. The reason for this, of course, cannot lie with the statistics. They are but tools in the possession of the "statistician." Like other "weapons of defense," they may be abused or misused. By themselves, they carry no significance. False conclusions are as easily supported by the use of statistics as are those which are true. One does not have to search widely for illustrations

[1] **McIlraith, James W., *The Course of Prices in New Zealand*, Government Printing Office, Wellington, New Zealand, 1911, p. 4 of *Introduction* by J. Hight.**

of this fact. For instance, in the hands of one, they are used to "prove" that railroad rates are too high; in those of another, that they are too low. As used by one, they seem to support the contention that wages have advanced; in those of another, that they have declined.

To what conditions are these different conclusions due? Motive in some instances; ignorance, in others. More often, however, they result because the following among other fundamental rules in the use of statistics are ignored:

"Never have preconceived ideas as to what the figures are to prove.

"Never reject a number that seems contrary to what you might expect, merely because it departs a good deal from the apparent average.

"Be careful to weigh and record *all* the possible causes of an event, and do not attribute to one what is really the result of a combination of several.

"Never compare data which have nothing in common."[1]

It is not our purpose at this place in the discussion to supply a set of rules for the use of statistics. As the treatment proceeds, this will be done in connection with the different topics discussed. It is, however, of interest to sketch briefly certain clearly marked tendencies by which beginners in the use of statistics and consumers of statistics are affected. Attention should be called to them in passing.

(1) The tendency to accept and to use without question any available "statistics." They are freely quoted, and cited at length when other methods fail. *Ipse dixit* is often regarded as sufficient proof. The mere fact that statistics are in print and appear in tabulated or graphic form—the finality of a statistical table, diagram or graph is often magical—serves to give them sufficient sanction. Of course, they may be inappropriate for the use to which they are put, and yet they are "statistics." Why not quote them when they are

[1] Newsholme, Arthur, *The Elements of Vital Statistics*, London, 1892, 3d Ed., pp. 292-293.

available, and when to the unsuspecting they carry profound weight? Illustrations of such tendencies are common. One has only to recall popular addresses, to consult the daily press, and to observe student reports in order to find examples of this practice. Teachers observe a kindred tendency in students to cite the statements from their textbooks as irrefutable proof. It is one part of the teacher's task to correct, and one portion of the student's training to overcome this tendency.

(2) The tendency to concentrate attention on statistical quantities or frequencies and to ignore the units in which they are measured. The *same* things or conditions are rarely counted for any length of time. Neither are the *same* units *of measurement generally used at different places. The uses* which statistics are intended to serve change from period to period. As a consequence, units of measurement also change. Moreover, different policies prompt statistical organizations at the same time but at different places to use different units, to interpret them in different ways, and to insist upon different standards of accuracy and completeness. These facts are frequently forgotten. But they ought not to be.

(3) The failure to remember that statistical compilations are generally made for definite purposes and that they cannot be used with the same precision for other purposes.

(4) The tendency to ignore the fact that statistics are in a very real sense personal. By this is meant the fact that some person or organization is responsible for them—that upon someone has been placed the responsibility of setting up the standards according to which they were collected, of determining upon the amount of error which would be tolerated, of mapping out the field from which they should be drawn, and of deciding upon the subjects to which they apply. But the personnel and policies of statistical organizations change, and with them also the continuity of statistical series.

(5) The tendencies to disregard detail—or to regard it as "detail" which somehow will take care of itself and needs no

especial attention; to ignore statistical cautions respecting the collection of data or the use of those already collected; to speak in terms of statistical abbreviations, averages of all types; to employ totals as if they were always more accurate than the items which go to make them up; and to piece together statistical fragments, gleaned from widely different sources and compiled under widely different circumstances and conditions.[1]

But to call attention to these tendencies is not sufficient to correct them. More is necessary. Students need to be shown the consequences to which they lead. Moreover, they must be instructed in what the scientific uses of statistics consist. It is one of the purposes of this volume to put the reader in possession of the information, tools, and knowledge whereby he can use and interpret statistics intelligently. Moreover, it is intended to supply information which will help him to pass upon the merits of the statistical approach to economic, social, and business problems, and to undertake statistical studies independently.

[1] For an admirable discussion of the false uses to which statistical data will be put, even by those who are in a position to know their limits, when it is a question of making a case, see Bowley, A. L., "Statistical Methods and the Fiscal Controversy" in *The Economic Journal*, London, Vol. 13, 1903, pp. 303-313. In formulating the rules to be observed, Bowley says:

"Every statistical estimate should be considered in the light given by corresponding estimates for previous years.

"Every total should be homogeneous in that quality which concerns the argument.

"Where values are used, the effect of replacing them by quantities should be tested.

"The errors latent in the constituents which form an estimate should be examined, and their effect on the estimates should be tested with reference to the purpose for which the estimate is used. The maximum adverse errors should be calculated, to see if their concurrence would vitiate the result.

"The ideal measurement necessary to support each deduction should be conceived; and if the estimates accessible do not necessarily give the same view as the ideal measurement, they should be rejected.

"When the sufficiency of statistics as estimates is established, the arguments based on them should be bound to the statistical results by the ordinary rules of logic." *Ibid.*, p. 312.

## II. The Meaning of Statistics and Statistical Methods

*Statistics* are generally thought of from two points of view: *first,* as series of numerical facts; and *second,* as methods which have to do with the collection, classification, tabulation, summation, abbreviation, and comparison of such facts for the purpose of describing or explaining the phenomena with which they deal. The first point of view is concerned with the finished product—the facts themselves; the second, with the preparation of the raw material and with the use of the finished product.

The two ways of looking at the subject are complementary. To secure the final product—statistics—requires the use of methods. These are concerned primarily with the technique of collection—enumeration and estimation—and with summation and abbreviation. The use of statistics—statistical methods—closely approaches logic, concerned as it is with the processes and methods of formulating and testing conclusions from premises which rest solely upon statistics. The conditions which determine what shall be enumerated; the units which shall be used; the accuracy, completeness, and consistency which shall be insisted upon, etc., largely determine the methods to be used in analysis. It is an error to think of the two viewpoints as unrelated. They are intimately connected. The adequacy of a tool, or the perfection of a machine—to speak analogously—is quite as important in the determination of a product as is the way in which it is used. Of course, skillful use may in part compensate for a poor tool, as skillful discrimination in the use of statistics may tend to correct errors following from crude or defective enumeration or estimation. An accurate statistical conclusion may sometimes be reached by the use of inaccurate data. But such is not the rule. Statistics, as methods, are as much concerned with the preparation of the final product—statistics—as with their use. In what follows, the principles of methodology are extended to both phases of the subject.

In definitions of statistics the emphasis has been variously placed. Bowley has called statistics the "science of averages"[1] as well as "the science of counting."[2] The first definition emphasizes one device for statistical abbreviation; the other calls attention to the enumeration which precedes analysis. In another place, Bowely defines *statistics* as "numerical statements of facts in any department of inquiry, placed in relation to each other," and *statistical methods* as "devices for abbreviating and classifying the statements and making clear the relations."[3] Yule defines *statistics* as "quantitative data affected to a marked extent by a multiplicity of causes" and *statistical methods* as "methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes."[4] Pearl defines statistics as "that branch of science which deals with the *frequency* of occurrence of different *kinds of things* or with the *frequency* of occurrence of different *attributes* of things."[5] Still others, using the terms with less precision, and in a less scientific sense, have sought to identify statistics with graphic methods—to convert the science into an art.

We shall use the term *statistics* as meaning *aggregates of facts, "affected to a marked extent by a multiplicity of causes," numerically expressed, enumerated, or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose, and placed in relation to each other.*

This definition needs to be explained. *Statistics are always aggregates:* that is, they are made up of a number of cases. *Isolated facts* are *not* statistics: they may be the instances

[1] Bowley, A. L., *Elements of Statistics*, P. S. King, London, 4th Ed., 1920, p. 7.

[2] *Ibid.*, p. 3.

[3] Bowley, A. L., *Elementary Manual of Statistics*, MacDonald & Evans, London, 1915, p. 1.

[4] Yule, G. U., *An Introduction to the Theory of Statistics*, Griffin & Company, London, 1911, p. 5.

[5] Pearl, Raymond, *Introduction to Medical Biometry and Statistics*, W. B. Saunders Company, Philadelphia, 1923, p. 19.

which make up statistics, provided they relate to the same thing over a period of time, to different attributes of things, or to the same thing at different places or times. A single death, an accident, a sale, a shipment does not constitute statistics. Yet numbers of deaths, accidents, sales, and shipments are statistics. Why? Because they are aggregates which may be analyzed: that is, studied in relation to time, place, and frequency of occurrence.

Moreover, *statistics are "affected to a marked extent by a multiplicity of causes."* They refer to measurements of phenomena in a complex universe. They are related to other measurements. They grow out of a variety of circumstances, differing among themselves, and are constantly subject to change. None of them are traceable to a single cause.

*Statistics, moreover, are numerically expressed.* Quantities not qualities are dealt with. Differences are shown by number. For instance, crops over a series of years, expressed in bushels harvested per acre, are statistics. The same facts indicated by such expressions as "good," "fair," "medium," "poor," etc., are not statistics unless a numerical equivalent is assigned to each qualitative expression.

*Statistics*, if they are to serve as the basis for a logical conclusion, and are to be combined, averaged, and summarized, *must be enumerated or estimated according to reasonable standards of accuracy.* Moreover, the same standards must obtain throughout the whole process of collection. What standards are "reasonable" depends upon the purpose which the statistics are to serve. No absolute criterion can be established for all cases. Where precision is required, accuracy is necessary; where general impressions are sufficient, appreciable error may be tolerated.

Then, too, if quantitative measurements are truly to be called "statistics," *they must be made in a systematic manner in keeping with a given purpose.* The purpose for which things are counted, or measurements and estimates made, will always determine the standards followed. If the purpose changes,

quantities may still be secured, but they refer to different things, or to the same thing in different ways, or to different degrees. They cannot be treated statistically and become the basis for valid conclusions.

For quantities to be called statistics, moreover, *they must be capable of being placed in relation to each other.* This may be done in point of time, of place, or of condition. That is, the term suggests comparison, and in order for things to be compared, they must have qualities in common. Indeed, as Bowley says, "Like can only be compared with like."[1] Stray and loose bits of quantitative information, hearsay, and unrelated material, gleaned here and there from indiscriminate sources, having no common basis of selection, while numerical, can be termed statistics only by a confusion of terms. If they are aggregates, homogeneous in the qualities necessary for comparison, then they may be called statistics, but not otherwise.

So much for the definition of *statistics.* But the term is used in another sense. It is sometimes spoken of as a science. In this usage, it refers to a method or to methods of dealing with the frequencies with which different things, or different attributes or characteristics of things occur. In some cases, it is spoken of as *a method;*[2] in others, *as methods.* We shall use the term in the plural.

*Statistical methods include all those devices of analysis and synthesis by means of which statistics are scientifically collected and used to explain or describe phenomena either in their individual or related capacities.*

[1] Bowley, A. L., "The Improvement of Official Statistics," in the *Journal of the Royal Statistical Society,* September, 1908, Vol. 71, p. 467.

This article is reprinted in the author's *Readings and Problems in Statistical Methods,* Macmillan & Company, New York, 1920, pp. 150-159.

[2] "The *statistical* method is that which deals with assemblages, or groups, in terms of the averages by which they may be described, and deals with relations which are not described by unchanging laws but by generalizations couched in terms of approximations and of probability." Mills, Frederick C., "On Measurement in Economics," in *The Trend of Economics,* Knopf, New York, 1924, pp. 38-39.

These methods have to do with the processes of (1) selecting and collecting data, (2) classifying them according to their common characteristics, (3) recording and illustrating the instances in keeping with a scheme of classification, (4) summarizing or abbreviating the detail by the use of averages, and (5) measuring the relationship which obtains between them. These are the methods with which the remaining part of this volume is concerned.

## III. The Use and Application of Statistics and Statistical Methods

Statistics are now collected on most important business and social problems. Indeed, we are surfeited with statistics. Some of them will satisfy our definition; others will not. This does not mean, however, that there is no dearth of statistics. There is. On many problems we have no adequate data. There is an abundance in some fields, and a scarcity in others. This condition is due to the growing need for information, part of which cannot be collected until plans have been developed. It is also due to the overlapping jurisdictions and conflicting purposes of public and private statistical organizations. Moreover, private purposes and transient needs prompt collections to be made, the series being discontinued as soon as the need is met, or changed in scope and meaning as soon as the purpose is served. The production of statistics is in a chaotic state; their use is hardly less haphazard.

But progress is being made. This extends *first* to their use. They *are* being employed, and this fact is significant. Discriminating use will come with an appreciation of their meaning to trade, industry, and the state, and with the development of skilled workers who know how to employ them. *Second,* progress is also being made in standardizing the methods of collection and presentation. Government departments, learned societies—such as The American Statistical Association—research organizations, etc., are all co-operating to improve and

extend not only the types of data collected but also to develop a technique of methodology in their use. The prospects are encouraging because statistical method is a "working tool of science. It is probably of wider utility than any other single tool which science has discovered or devised. For it has an applicability and a usefulness, direct or indirect, in virtually every problem. It is, in short, a fundamental element of scientific methodology."[1]

And yet, it is but one method. There are others which are often helpful in the explanation of phenomena. It has its limitations. It takes account only of quantitative and not of qualitative differences. It is not of universal use or validity. Yet when other methods are employed, statistics may often be used in a corroborative way. Indeed, it is in this respect that they probably have their greatest value.

This, however, does not mean that the function of statistics is limited to particular kinds of questions. There are few problems relating to business, social policy, or statecraft for an understanding of which statistics are not required. There is need everywhere for an appreciation, measurement, and analysis of facts in their quantitative aspects, for the ability accurately to observe the conditions to which they are traceable, for a determination logically and scientifically to piece them together, so that from them conclusions can be drawn which will become the basis for a program looking toward economic and social progress.

The fields of application of statistics and statistical methods, even to problems of economics and business alone, are too broad and varied to be described at this place. Some of them have already been mentioned. It may be helpful, however, to enumerate the *types* of problems which may be statistically studied. The subsequent discussion and illustrations will serve more definitely to develop the precise manner in which they may be and are being studied.

[1] **Pearl, Raymond, *op. cit.*, p. 21.**

1. *Application to Individual Business Units. A study of:*
   (1) Prices.
   (2) Production by departments, processes.
   (3) Sales and sales possibilities by districts, by periods, by products.
   (4) Employment, as to rapidity of turnover, scale of wages, labor supply, types of welfare work.
   (5) Factory organization and stock control.
   (6) Margins on different goods.
   (7) Costs; results of management policies; avenues of distribution; advertising methods and results; layout; price policies; trade practices; consumer-demand; credit risks; size, frequency, etc., of customer-purchases.
   (8) Profits—gross and net—by periods, by departments, by products.

2. *Application to Groups of Business Units.* Studies of this character might extend, among other things, to comparisons of:
   (1) Production. These would include:
      a. amounts and proportions of land, labor, and capital.
      b. expenses incurred and their distribution.
      c. materials used—sources, amounts, costs, shipments, storage, inventories, purchases.
      d. output—amounts, types, costs, distribution.
   (2) Finances:
      a. prices.
      b. capital requirements, source, kinds.
      c. relation of current assets to current liabilities.
   (3) Expenses:
      a. overhead, current, selling.
      b. relation of each expense to sales and to total expenses.
   (4) Margins:
      a. on different goods.
      b. in relation to sales.
   (5) Turnover of
      a. merchandise, by lines.
      b. capital.
      c. accounts receivable.
      d. inventories.
   (6) Profits—gross and net. Relation to
      a. total capital.
      b. sales.
      c. net worth.

3. Application to Matters of General Business Growth, Decline and Change. Under this head fall such topics as the following:
   (1) Production.
      a. production—value, quantities, and grades.
      b. stocks of goods—in sight, and potentially available.
      c. shipments.
      d. consumption.
   (2) Prices, money, and credit.
      a. banking activity—loans, discounts, debits, clearings.
      b. credit—interest rates, security issues and prices.
      c. security markets.
   (3) Labor supply and compensation.
      a. employment and unemployment.
      b. immigration, emigration, labor turnover, wage rates.
   (4) Economic waste of
      a. materials.
      b. human resources.
      c. transportation.
   (5) Characteristic features and sequence of economic factors during periods of
      a. prosperity.
      b. liquidation.
      c. stagnation.
      d. recovery.

4. Application to Questions of Social Economy.
   (1) Poverty, crime, dependency.
   (2) Consumption of goods and spending of incomes.
   (3) Growth, decline, and movements of population.
   (4) Mortality, sickness, accidents.
   (5) Occupational distribution and adjustments.
   (6) Farm and home ownership, tenancy.
   (7) Distribution of wealth and income.
   (8) Conservation of natural resources.
   (9) Methods of wholesale and retail distribution.
   (10) Public expenditures, debt, taxes.

5. Application to Affairs Pertaining to Governmental Discrimination and Policy.
   (1) The determination of the benevolent or malevolent effects of given state policies, such as those pertaining to tariff, use of natural resources, price fixing, public ownership and control.

(2) The determination of "fair values" and "reasonable returns" as bases for the exercise of administrative discrimination and the shaping of governmental policy.

(3) The supervision of private business methods, looking toward the insuring of competition, the regulation of monopoly, the guaranteeing of favorable conditions of employment.

(4) The evaluation of properties as a basis for taxation, condemnation, and forced sale.

(5) The recording of domestic and foreign trade movements, estimating national wealth and its distribution, recording national progress so far as revealed statistically.

6. Application to Questions of Economic Theory.

The science of economics is becoming statistical in its method.[1] The advice of Richard Jones to "Look and see" is being taken literally. Accordingly, in the study of the law of demand, for instance, recourse is being made to statistics of markets where demand is indicated in the prices paid and amounts purchased. Similarly, supply is studied with respect to costs, these being measured in standard units. Market analyses and cost studies are now becoming commonplaces, albeit that they are for the most part undertaken only by the larger business units, and are far too often unscientifically carried out. The significant thing is that they are being made. Improvement will come in time. Just as fast as business men, singly or in groups, come to realize that there are basic principles which lie behind the daily routine of pricing, producing, and selling, for instance, which may be discovered and stated, just so fast will they seek for and be guided by such principles.

Jevons, in 1871, stated the problem clearly. He said, "I know not when we shall have a perfect system of statistics, but the want of it is the only insuperable obstacle in the way of making Economics an exact science."[2] Keynes says that

[1] Tugwell (Editor), *The Trend of Economics*, Alfred Knopf, New York, 1924, Chapters I and II, pp. 3-34, and 37-70, respectively.

[2] Jevons, W. Stanley, *The Theory of Political Economy*, Macmillan & Company, New York, 4th Ed., 1911, p. 12.

the function of statistics is "first, to suggest empirical laws, which may or may not be capable of subsequent deductive explanation; and secondly, to supplement deductive reasoning by checking its results, and submitting them to the test of experience."[1] Professor Moore's *Laws of Wages* is an excellent example of the use of statistics and statistical methods in the development of economic theory. Stating his purpose, he says, "I have endeavored to use the newer statistical methods and the more recent economic theory to extract, from data relating to wages, either new truth or else truth in such new form as will admit of its being brought into fruitful relation with the generalizations of economic science."[2]

The use of statistics and statistical methods for these purposes, while possessing great possibilities in the hands of the well-trained statistical economist, offers few opportunities to the readers to whom this volume is addressed.[3]

[1] Keynes, J. N., *Scope and Method of Political Economy*, 2d Ed., revised, Macmillan & Co., London, 1897, p. 338.

[2] Moore, H. L., *Laws of Wages*, Macmillan & Company, New York, 1911, p. 6.

[3] It may be of general interest to list some of the economic subjects with respect to which statistics have been used to discover "laws" or tendencies. Among these are the following: the business cycle, competition, consumption, distribution of wealth and income, population growth, prices, production, rents, trade, unemployment, wages, etc. There is an extensive literature pertaining to these subjects. Those who are interested may consult the following among other writings:

ON THE BUSINESS CYCLE

Hansen, Alvin H., *Cycles of Prosperity and Depression in the United States, Great Britain, and Germany—A Study of Monthly Data, 1902-1908*, Madison, Wisconsin, 1921.

*Business Cycles and Unemployment*, McGraw-Hill, New York, 1923.

Mitchell, Wesley C., *Business Cycles*, Univ. of California, Berkeley, 1913.

Moore, H. L., *Economic Cycles, Their Law and Cause*, Macmillan & Company, New York, 1914.

Moore, H. L., *Generating Economic Cycles*, Macmillan & Company, New York, 1923.

Moore, H. L., *Forecasting the Yield and the Price of Cotton*, Macmillan & Company, New York, 1917.

Persons, W. M., "The Construction of a Business Barometer based upon Annual Data" in *American Economic Review*, December, 1916, pp. 739-769.

With this introduction, the purpose of which is to open up the subject, to define its boundaries, and to suggest the nature of the uses of statistics and statistical methods, we pass immediately, in Chapter II, to a consideration of *Types of Secondary Statistical Data and Tests for their Use.*

(Note 3 continued)

PERSONS, FOSTER and HETTINGER (Editors),, *The Problem of Business Forecasting*, Houghton Mifflin, Boston, 1924, *passim.*

*Review of Economic Statistics*, Harvard Economic Service, Cambridge, Mass., especially the numbers for January and April, 1919; July, 1923; January, 1924.

ON COMPETITION, COSTS, DEMAND, AND PROFITS

SCHULTZ, HENRY, "The Statistical Measurement of the Elasticity of Demand for Beef," *Journal of Farm Economics*, July, 1924, pp. 254-278.

SECRIST, HORACE, "Competition in the Retail Distribution of Clothing—A Study of Expense or 'Supply' Curves," *Bureau of Business Research, Northwestern University*, Chicago, 1923.

SECRIST, HORACE, "Expense Levels in Retailing—a Study of the 'Representative Firm' and of 'Bulk-Line' Costs in the Distribution of Clothing," *Bureau of Business Research*, Northwestern University, Chicago, 1924.

SIMPSON, KEMPER, "A Statistical Analysis of the Relation between Cost and Price," *Quarterly Journal of Economics*, 1921, pp. 264-287.

SIMPSON, KEMPER, "Further Evidence on the Relation between Price, Cost, and Profit," *Quarterly Journal of Economics*, February, 1923, pp. 476-490.

TAUSSIG, F. W., "Price Fixing as Seen by a Price Fixer," *Quarterly Journal of Economics*, February, 1919, pp. 205-241.

WRIGHT, PHILIP G., "Value Theories Applied to the Sugar Industry," *Quarterly Journal of Economics*, November, 1917, pp. 101-121.

WRIGHT, PHILIP G., *Sugar in Relation to the Tariff*, McGraw-Hill, New York, 1924, pp. 106-130; 276-284.

ON CONSUMPTION

OGBURN, W. F., "Analysis of the Standard of Living in the District of Columbia in 1916," in *Quarterly Publications of the American Statistical Association*, June, 1919, pp. 374-392.

ON DISTRIBUTION OF WEALTH AND INCOME

*Income in the United States—Its Amount and Distribution*, 1909-1919. National Bureau of Economic Research, New York, Vol. I, 1921, Vol. II, 1922.

ON POPULATION GROWTH

PEARL, RAYMOND, and REED, LOWELL, J., *Predicted Growth of the Population of New York and its Environs*, New York, 1923.

ON PRICES

FISHER, IRVING, *The Purchasing Power of Money*, Macmillan & Company, New York, 1911.

### References

BOUCKE, O. FRED, *A Critique of Economics*, Macmillan & Co., New York, 1922, pp. 211-231.

BOWLEY, A. L., *Elements of Statistics*, P. S. King & Son, London, 1911, Chapter I, pp. 3-13.

BOWLEY, A. L., *An Elementary Manual of Statistics*, MacDonald & Evans, London, 1915, Chapter I, pp. 1-6.

JONES, D. CARADOG, *A First Course in Statistics*, C. Bell & Sons, London, 1921, Chapter I, pp. 1-4.

KING, W. I., *Elements of Statistical Method*, Macmillan & Co., New York, 1912, Chapters II, III, pp. 20-39.

KEYNES, J. M., *A Treatise on Probability*, Macmillan & Co., London, 1921, Chapter XXVII, pp. 327-331.

MILLS, FREDERICK C., "On Measurement in Economics," in *The Trend in Economics*, Alfred Knopf, New York, 1924, pp. 37-70.

PEARL, RAYMOND, *Introduction to Medical Biometry and Statistics*, W. B. Saunders, Philadelphia, 1923, Chapter I, pp. 17-25.

PEARSON, KARL, *The Grammar of Science*, 3rd Edition, Adam and Charles Black, London, 1911, Chapter I (Introduction), pp. 1-38.

RUGG, HAROLD O., *Statistical Methods Applied to Education*, Houghton Mifflin, Boston, 1917, Chapter I, pp. 1-27.

SECRIST, HORACE, *Readings and Problems in Statistical Methods*, Macmillan & Company, New York, 1920, Selections in Chapter I, pp. 1-46.

WEST, CARL J., "The Value to Economics of Formal Statistical Methods," in *Publications of the American Statistical Association*, September, 1916, pp. 618-628.

YULE, G. UDNY, *An Introduction to the Theory of Statistics*, Griffin & Company, London, 1911, Introduction, pp. 1-6.

(Note 3 continued)

FISHER, IRVING, *The Making of Index Numbers*, Houghton Mifflin, Boston, 1923.

KEMMERER, E. W., *Money and Credit in their Relation to General Prices*, Holt, New York, 1907.

MITCHELL, W. C., "History of Prices During the War," *War Industries Board*, Washington, D. C., 1919.

ON PRODUCTION INDEXES

DAY, E. E., "An Index of the Physical Volume of Production," *Review of Economic Statistics*, Harvard Economic Service, Cambridge, Mass., September, 1920, to January, 1921.

KING, W. I., *Bankers' Statistics Corporation*, Special Service, New York, Vol. II, No. 12, August 24, 1920.

(Note 3 continued)

LEONARD, W. E., "*An Index of Changes in Extractive Industries*," *Quarterly Publications American Statistical Association*, September, 1913, pp. 539-550.

STEWART, W. W., "An Index Number of Production," *American Economic Review*, Vol. XI, No. 1, March, 1921, pp. 57-81.

ON RENTS

SECRIST, HORACE, "Commercial Rent as an Expense and its Relation to Profits," *Bureau of Business Research*, Northwestern University, Chicago, 1923.

ON TRADE

PERSONS, W. M., "An Index of Trade for the United States," in *Review of Economic Statistics*, Harvard Economic Service, Cambridge, Mass., April, 1923, pp. 71-78.

SNYDER, CARL, "*A New Index of the Volume of Trade*," *Journal of the American Statistical Association*, December, 1923, pp. 949-963.

ON UNEMPLOYMENT AND EMPLOYMENT

BERRIDGE, WM. A., *Cycles of Unemployment in the United States*, 1903-1922, Boston, 1923.

KING, W. I., *Employment Hours and Earnings in Prosperity and Depression*, 1920-22, National Bureau of Economic Research, 1923.

ON WAGES

MOORE, H. L., *Laws of Wages*, Macmillan & Company, New York, 1911.

# CHAPTER II

## TYPES OF SECONDARY STATISTICAL DATA AND TESTS FOR THEIR USE

### I. Introduction

For statistics to be used, they must be available. Indeed, the way in which they are used is determined by the conditions which have been or may be followed in collecting or assembling them. Statistics do not come into being of and by themselves. They are not collected without a purpose. Those which are now available were originally intended to serve some end, notwithstanding the fact that it may not be apparent to the user and may be foreign to the needs of a particular time, place, or condition. This must not be forgotten. Likewise, those which are in process of collection, or are to be collected, will be chosen because of their suitability to a definite purpose.

At any given time or place, or under any condition, the collection of statistical data presupposes certain standards of accuracy, completeness, and comparability. What these are for any group of data depend upon (1) the purpose in mind, (2) the character of the data themselves, (3) the bases of selection and omission, (4) the integrity, honesty, and organization of the collecting body, (5) the basis of classification used in grouping them, (6) the clerical accuracy used in their compilation, and (7) the adherence to uniform units or terms in which the quantities are expressed.

Statistics are found either as a "finished product" or as "raw material." In the first form, they appear in the trade press, government documents, newspapers, annual reports of banks,

corporations, etc.; in the second form, in the transactions of business, the processes of industry, movements of population, etc. They are constantly in a state of "manufacture." The finished product of to-day is the raw material of yesterday.

This chapter has to do: *first*, with a brief description of the chief sources of secondary statistical data, that is, with those already available; and *second*, with the tests which should be applied to such data before they are used.

It is not our intention to furnish an exhaustive list of the different types of secondary statistical data, nor to indicate all of the places where they may be found. Neither shall we attempt to give a complete list of the private and public organizations which collect such data. The present output of statistics is enormous. It applies to a vast and constantly changing number of subjects, and is of different value at the same time at different places, and at different times at the same place. To write a critique of secondary data would be an extremely difficult if not impossible task. Moreover, it would be of little permanent value, since the methods which govern their collection change from time to time in the light of the particular needs and standards of the bodies responsible for them. This much, however, may be said: the value of the output is improving; statistical organizations, both public and private, are being placed on a substantial and permanent basis; and statistical data, because of the use to which they are put, are being subjected to critical tests. These have to do, among other things, with completeness, accuracy, and uniformity. But more concerning them presently.

Statistics, as indicated above, are numerical aggregates having certain well-defined properties. They are syntheses [1] made

[1] "When we are investigating the nature and causes of things and events in the natural and social sciences, we are face to face with *facts*. In statistics about those events we are brought face to face with *syntheses*. The statistician must regard his figures as a sort of symbol, whose character and significance are more or less enigmatic; and he must diligently seek out all the probable causes of the facts he has symbolized before him, with a view to their scientific explanation." P. Coffey, *The Science of Logic*, Longmans, London, 1912, Vol. II, p. 287.

up of individual instances. Moreover, they are derivative in the sense that they numerically measure phenomena as they appear to an observer. The identity of the parts of even the simplest statistical aggregate must be established. Identification requires that "earmarks" shall be distinguished, and that they shall always appeal in the same way to those who are responsible for making the selection. To count such simple things as bushels of wheat, for instance, appears to be easy. Yet it is not always clear what is meant by a "bushel," nor what is included in the term "wheat."[1] Similar observations may be made about any statistical data. The important points to be considered are: (1) what are counted, (2) are the same things always included, (3) who did the counting, and (4) for what purpose was the counting made? These topics need to be more fully considered. The discussion for the moment, however, has to do with the distinction between primary and secondary data. It will later include (1) the chief sources of secondary data, and (2) the tests to which such data should always be subjected before they are used.

## II. Primary and Secondary Data Defined and Contrasted

It is necessary to define, more accurately than has been done above, what is meant by secondary data. By "secondary data" are meant those which have been collected, tabulated, and presented in simple or complex form for any purpose whatsoever. They generally appear as totals or percentages, removed one or more steps from the form in which they were reported. Consequently, they do not show on their face (1) the peculiarities of the units employed, (2) the purpose or purposes for which collected and used, (3) the way in which they have been edited, combined, and grouped, nor (4) the adjust-

[1] See the interesting study by Boerner, E. G., "Improved Apparatus for Determining the Test Weight of Grain, with a Standard Method of Making the Test," *Bulletin No. 472, U. S. Department of Agriculture, October, 1916.*

RESEARCH INST

ments which have been made in the original data in order that they might be used for the purpose in mind. They are truly "secondary." They have been carried through certain manipulations, the extent and character of which are not generally disclosed.

In contrast with such data are those which are called *primary.* By "primary data" are meant those which are original: that is, those in which little or no grouping has been made, the instances being recorded or itemized as encountered. They are essentially raw material. They *may be* combined, totaled, and averaged; but they *have not* extensively been so treated.

Of course, the distinction between primary and secondary data is largely one of degree. Data which are secondary in the hands of one party may be primary in the hands of another. Illustrations will make this clear. To the Federal Reserve Bank of Chicago, for instance, the reported debits to individual accounts of the member banks are primary data. To one reading the report of the bank showing the total debits for the district, they are secondary. To the general public, the death rates published by the Board of Health of Chicago constitute secondary data. In the hands of the statistician of this Board, they are primary data. Moreover, to the Bureau of Business Research, Northwestern University, the records of sales, expenses, inventories, etc., secured from the books of retail meat establishments, are primary data. When these same facts are published by the Bureau, in interpretive studies, they become secondary. Wherein lies the distinction? Essentially, in the fact that the data before publication have been edited for completeness, accuracy, comparability, consistency; they have been combined into groups, averaged, summarized, expressed as percentages, etc. They have been "worked over" for a purpose; they have lost the individual characteristics which they possessed as primary data when reported.

But even so-called "primary data" are in reality secondary to the degree to which they have been "worked over" in the

process of gathering. While the distinction between the two is largely one of degree, it is none the less important. It is significant because *the more secondary data become, the more specialized is their function, and the more difficult is it to use them for purposes other than those for which they have already been used.* Each successive use is made for a purpose, and carries with it new and different bases for combinations, adjustments, omissions, etc.

### III. Sources of Secondary Statistical Data

The chief sources of secondary statistical data are the periodic and occasional reports of (1) national, state, and city departments, bureaus, and commissions, (2) trade associations and private organizations, (3) research agencies, (4) technical periodicals.[1] Space is available for listing only a few of the representative sources falling under each of these headings, and for indicating the scope of the statistical material issued.

It is not in keeping with our purpose to compile a catalog of statistical sources, neither is it to our interest to make a compilation of the statistical material which is or might be of interest to students of business and social affairs. A certain amount of the foraging or exploring instinct, and at least a general knowledge of what data are likely to be available in the sources to which reference is made, are presupposed on the part of the person who has occasion to use published statistics. If such knowledge is lacking, it may be easily acquired by those who really seek it.

But it is inadequate alone to know the sources of statistical data. More is needed. The ability to pass judgment on the

[1] For a list of the main agencies, both public and private, together with a description of the nature of the data published, the name of the publication in which they are contained, and the date of publication, see *Survey of Current Business*, Monthly Supplement to Commerce Reports, United States Department of Commerce. This *Survey*, published monthly by the United States Department of Commerce, contains a selected body of data on matters pertaining to business.

value of such data is also necessary. In addition to both, training is required in the scientific use of the data for the purposes desired. It is primarily the last aspect of the problem in which our interest lies.

### A List of Some of the More Important Sources of Secondary Statistical Data

The Federal Government

- U. S. Department of Agriculture
  - Bureau of Agricultural Economics
  - Bureau of Animal Industry
  - Forest Service
- U. S. Department of Commerce
  - Bureau of the Census
  - Bureau of Foreign and Domestic Commerce
  - Bureau of Navigation
- U. S. Department of the Interior
  - Bureau of Mines
  - Geological Survey
- U. S. Department of Labor
  - Bureau of Immigration
  - Bureau of Labor Statistics
- U. S. Treasury Department
- Federal Reserve Board
- Federal Trade Commission
- Interstate Commerce Commission

The State Governments

- Illinois Department of Labor, Springfield
- Massachusetts Department of Labor and Industries, Boston
- New York State Department of Labor, Albany
- Pennsylvania Department of Labor and Industry, Harrisburg
- Wisconsin Industrial Commission, Madison
- Wisconsin Tax Commission, Madison

Research Agencies

- University
  - Brown University, Bureau of Business Research, Providence, R. I.
  - Carnegie Institute of Technology, Dept. of Commercial Engineering, Pittsburgh, Pa.

Harvard University, Bureau of Business Research, Cambridge, Mass.

New York State College of Agriculture, Cornell University, Department of Agricultural Economics and Farm Management, Ithaca, N. Y.

New York University, Bureau of Business Research, New York, N. Y.

Northwestern University, Bureau of Business Research, Chicago, Ill.

University of Colorado, Bureau of Business and Governmental Research, Boulder, Colorado

University of Illinois, Bureau of Business Research, Urbana, Ill.

University of Nebraska, Committee on Business Research, Lincoln, Neb.

University of Oregon, Bureau of Business Research, Eugene, Oregon

University of Pennsylvania, Industrial Research Department, Wharton School of Finance and Commerce, Philadelphia, Pa.

Other

American Institute of Agriculture, Chicago, Ill.
Bureau of Railway Economics, Washington, D. C.
Food Research Institute, Stanford University, California
Institute for the Study of Land Economics, Madison, Wis.
Institute of Economics, Washington, D. C.
International Institute of Economics, New York, N. Y.
Life Insurance Sales Research Bureau, New York, N. Y.
National Bureau of Economic Research, New York, N. Y.
National Industrial Conference Board, New York, N. Y.
Russell Sage Foundation, New York, N. Y.

Trade Associations and Private Organizations

American Face Brick Association, Chicago, Ill.
American Newspaper Publishers' Association, New York, N. Y.
American Iron and Steel Institute, New York, N. Y.
American Railway Association New York, N. Y.
Automobile Manufacturers' Association, Chicago, Ill.
Chicago Board of Trade, Chicago, Ill.
F. W. Dodge Corporation, Boston, Mass.
National Association of Farm Equipment Manufacturers, Chicago, Ill.

National Automobile Chamber of Commerce, New York, N. Y.
New York Coffee and Sugar Exchange, New York, N. Y.
Portland Cement Association, Chicago, Ill.
Silk Association of America, New York, N. Y.
United Typothetæ of America, Chicago, Ill.

This list of sources of secondary data refers to statistics of interest primarily to the business man and student of business. It is not intended to be complete. Reference should also be made to the matter contained in the footnote below.[1]

These sources contain statistical data of the "secondary" sort. To pass judgment upon their merits even for a specific purpose would involve an enormous amount of study and discrimination, since each collection has its own peculiarities and is collected with a given end in view. To judge of their value

[1] For an account of the sources of statistics on produce markets, see Mudgett, Bruce D., "Current Sources of Information in Produce Markets," in *Annals of the American Academy of Political and Social Science*, Vol. XXXVIII, No. 2, pp. 104-125. On some of the private organizations regularly collecting and issuing statistical data, see Parmelee, Julius H., "The Utilization of Statistics in Business," in *Quarterly Publications of the American Statistical Association*, June, 1917, pp. 565-576. See also Haney, Lewis H. and Meyer, C. C., *Source Book of Research Data*, Prentice-Hall, New York, 1923; West, Carl J., *Market Statistics*, U. S. Department of Agriculture, Washington, D. C., Bulletin 982, June, 1921; *Statistical Abstract*, Department of Commerce, Washington, D. C.

The student who has or wishes to cultivate an interest in statistics pertaining to business should regularly consult the following, among other, publications:

*The Federal Reserve Bulletin*, The Federal Reserve Board, Washington, D.C.
*The Monthly Reviews of Business Conditions*, The Respective Federal Reserve Banks.
*The Monthly Labor Review*, U. S. Department of Labor, Washington, D.C.
*The Review of Economic Statistics*, Harvard Committee on Economic Research, Cambridge, Mass.
*Harvard Economic Service*, Harvard Committee on Economic Research, Cambridge, Mass.
*The Brookmire Economic Service*, New York, N. Y.
*Babson Statistical Service*, Wellesley Hills, Mass.
*Dun's Review*, R. G. Dun, New York, N. Y.
*Bradstreet's*, The Bradstreet Company, New York, N. Y.
*The Annalist*, New York, N. Y.
*Moody's Investors Service*, New York, N. Y.
*Commercial and Financial Chronicle*, Wm. B. Dana, New York, N. Y.
*The Journal of the American Statistical Association*, Columbia University, New York, N. Y.

for *general* purposes is impossible, because no criteria of distinction are offered. Yet, it is not impossible to point out certain tests to which they should all be subjected before they are used. It is the purpose of the following section to outline such a series of tests.

## IV. Tests to be Applied to Secondary Statistical Data Before They are Used

The inquiries which should always be made about secondary data relate to (1) the organization which supplies the data, (2) the purpose for which they are issued and the consumers to whom they are addressed, (3) the nature of the data themselves, (4) the units in which expressed, (5) their accuracy, (6) the extent to which they refer to homogeneous conditions, and (7) their application to a given problem. Each of these topics requires special consideration.

### 1. THE ORGANIZATION SUPPLYING SECONDARY DATA

Every statistical organization is created for a purpose and has a special function to perform. Some are public, some semi-public, and others private. Some are old and have well-established standards of excellence; others are relatively new—are struggling to secure information, and trying to present it in a form suitable to a special clientèle. Some are adequately financed and have proper entrée to sources of information; others are financially embarrassed and must be content to secure information from any source available. Some have legal sanctions to compel information to be furnished in keeping with a carefully prepared plan relating to each detail covered; others must be content with information gratuitously furnished, and in a form which suits the interest, prejudice, or peculiar records of informants.

If these and other differences characterize organizations which publish statistical data, then the person who has occasion to use such material must ask, and answer to his own satisfaction, the following, among other, questions:

(1) What types of organizations issue the data desired?
(2) Is there a choice between them?
(3) What standards of excellence obtain in their collection, and in their interpretation?
(4) Is there anything in the nature of the organization which might prejudice the data in any vital particular?

Some information about all of these inquiries is available. It may be difficult to secure, and be incomplete, yet, to any one who really desires it, methods are available by which it may be secured. Any responsible statistical organization is glad to describe its form of organization and its methods.

### 2. THE PURPOSE FOR WHICH SECONDARY DATA ARE ISSUED AND THE CONSUMERS TO WHOM THEY ARE ADDRESSED

Whatever may be the type of its organization, each statistical body has its own policy and its particular purpose. Accordingly, there is generally some basis for a choice between sources, notwithstanding the fact that they *appear* to present the same or similar data, and to serve the same clientèle. Choice will generally depend more upon the purpose which an organization serves than the type of the organization itself. These purposes may be:

(1) General or specific
(2) Restrictive or inclusive
(3) Transient or permanent
(4) Scientific or unscientific

Because of these differences in the purposes for which data are collected and published, secondary data ought not to be used indiscriminately. They are good or bad, satisfactory or unsatisfactory, in the light of the purpose which controlled their collection or selection, their grouping and combination, and the analysis which has been made of them.

### 3. THE NATURE OF THE SECONDARY DATA THEMSELVES

In the use of secondary data, after the type of organization which issues them and the purposes which they are intended to

serve have been determined, the data themselves must be examined. The following among other facts should be considered:

(1) Are the data biased? Bias may be due to (a) wilfully eliminating parts of the facts, (b) basing comparisons upon insufficient data, or (c) relating them to unrepresentative periods or conditions. When prompted by motives to deceive, little difficulty is found in making out a case from data which if otherwise used would tell a different story. If samples are chosen according to chance, an accurate account may be secured from comparatively few data. If, on the other hand, choice is biased, the effect of increasing the number of samples serves to increase the amount of error. No use should be made of secondary data until the question of bias is settled.

(2) Are the data samples only, relating to (a) restricted groups or characteristics, (b) certain territories, (c) particular times; or are they complete for the subject matter to which they relate?

Are all instances or frequencies included, or are samples selected: that is, are data inclusive or exclusive? Samples, in the very nature of the case, are generally used. The entire "population"—that is, *all* of the instances—save in studies based upon counts, are rarely included. Sampling, moreover, has to do with given times, classes or characteristics, and places. What bases of selection have been employed? How nearly do the samples describe the conditions to which they relate? *A satisfactory sample must contain the characteristics common to the entire "population," and these must be represented in the same proportions as they are found in the material sampled.*

If data constitute a census, then they must be complete. Instances or cases, no matter how typical of a group or class, cannot be omitted. By hypothesis, they must be complete. If, however, they are taken as representative of a class, then comparatively few instances may suffice for a sample, provided they are chosen at random, or with intent to in-

clude in suitable proportions the characteristics of the whole. Illustrations of problems requiring all of the data available, and of others which may be studied from samples, may help to make the discussion clear.

The total population of the United States cannot be known without the inclusion of every one; the sex composition may be accurately determined from a well-selected sample. Similarly, the total retail sales of meat products in Chicago cannot be known if the sales of a single merchant are excluded. The (average) cost of selling meat, however, may be accurately known from the records of an adequate sample. Again, if one were interested in the question of farm ownership and tenancy in a state, for instance, it would probably be necessary to study more than widely scattered sections, since conditions are not necessarily homogeneous as to the prevalence of ownership, nor uniform respecting the terms under which tenancy exists. If the types, amounts, and economic status of immigrant labor in the United States were being studied, one would hardly be safe in using data for a single state or city. It might be possible by so doing to secure data which are typical of the total immigration, but more than typical facts are wanted. The problem suggests a quantitative and not alone a qualitative result. The same is true respecting studies of births, deaths, accidents, etc. To record an occasional death, birth, or a few of the serious industrial accidents is inadequate. It is necessary to include all deaths, all births, and all accidents. Accident risks, for instance, cannot be properly determined unless all accidents occurring, the place where and the condition under which they happen, and the extent of disability, etc., are known.

On the other hand, if all that is desired is to indicate the trend in a given set of facts, it may suffice to take well-distributed samples. Changes in prices can be statistically determined without including statistics of all prices. The movement of wholesale prices, over a period of time, can be measured by using the prices of a comparatively few well-selected

commodities. The same is true of price changes of raw products, or of goods in which the final consumer is interested. The trend of the price of real estate, or of stocks and bonds, may be measured by the use of comparatively few but representative sales. Wage increases or decreases may be shown by a process of sampling, provided the samples are chosen with discrimination. An illustration of a case where samples suffice is found in the use by real estate boards and tax bodies of sales statistics in order to determine either the "market" or "true value" of real estate. The chief consideration is the representative character of the samples.

If it is desired, for instance, as evidence of the value of a piece of property, to enumerate the number of people who pass it, it is sufficient to include relatively short periods typical of both rush and slack hours for representative days. Likewise, the scale of rents in a given district may be determined with sufficient accuracy for commercial purposes by considering rents of representative houses. It is not necessary to include all houses rented. Care must always be exercised, however, to see that the sampling, howsoever carefully made for purposes of original compilation, is suitable for the purposes in mind. It may be stated, as a general rule, that *the more nearly all data are included, the less is the likelihood of bias controlling, and the more readily can they be converted to a particular use.* Under such circumstances the particular facts desired may be more easily chosen and extraneous ones eliminated. Again, however, nothing better than general principles can be laid down as a guide to the appropriate use of secondary material. Discrimination and caution are essential in scientific study and in the formulation of valid conclusions.

But how is it to be known from secondary data, as published, what bases have been used in selecting the samples? The regrettable truth is, that in too many cases it cannot be known. Publications have a practice of omitting all qualifying statements; of removing from the tabulated data all explanatory details; and of expecting the reader to take on faith

the accuracy, completeness, and representativeness of the material which is published. Not infrequently one is at a loss to know anything about such data. Sources are not given, irreconcilable totals are not explained, and inconsistencies abound. Under such circumstances, "Discretion is the better part of valor." The student may better refuse to use data than to be continually in doubt as to their meaning, scope and significance.

### 4. IN WHAT TYPES OF UNITS ARE THE DATA EXPRESSED? ARE THEY THE SAME AT DIFFERENT TIMES, AT DIFFERENT PLACES, AND FOR ALL CASES AT THE SAME TIME OR PLACE?

Secondary data are always presented in units of time, of place, or of condition. They are given, for instance, by months, by districts, and by age or size groups. Are the "months" always of the same length, and do they always begin and end at the same time? Similarly, are the "districts" always of the same size and do they have the same boundaries? Again, are the age or size groups the same from "month" to "month" and from "district" to "district"? Do the "same" data in two publications refer to the same time, place, and condition? Can the material from one source be combined with or used in the place of that from another?

Moreover, are the same things counted from time to time, and from place to place? In what kinds of units are they expressed, and what criteria are used to distinguish them? What, for instance, is a commercial failure, a bank loan, a farm, etc., as published in compilations of statistical data? Are "failures" and "farms" always identified in the same way? If they are not, and the differences are unknown, then how valuable for comparative purposes are the data concerning them?

The units in which data are *expressed* are of three general types. For convenience, they may be classified as *simple* units, as *composite* units, and as *coefficients* or *ratios.*

By *simple* units are meant those in which one determining

consideration is prescribed. The ideas conveyed are general; classes only being distinguished. Most statistics of enumeration employ simple units: as, for instance, when persons, animals, acres, buildings, passengers, stocks, deaths, laws, sales, etc., are counted. In statistics of this type the disturbing elements due to inaccuracies in the units are reduced to a minimum. Nothing, of course, is said about the accuracy with which the units are defined, of the care with which the definitions are followed, nor of the accuracy with which the enumerations are made. The characteristic feature of such units is the presence of a *single* determining condition. This normally guarantees against the presence of as great, or of a greater degree of error than would be associated with conditions when units are *composite* in character. Such a unit as a "farm" might be easily defined and the statistics of farm be readily understood. When, however, the expression "improved" is added to this unit and it becomes composite, the scope of the definition and its application are restricted. Error may enter into it with the same readiness as into the other portion of the combined unit. Likewise, in statistics of "daily wages" or of a "fair return," the same observation applies. Crops in bushels or in acreage may be readily determined—whether those crops are "normal," however, raises further questions. As limiting conditions are added to simple units, occasions for error and bias crowd in, and it is these to which attention is drawn in distinguishing simple from composite units.

Statistical data may also be expressed as *ratios* or *coefficients*. The units then take the form of comparative statements: as, for instance, when deaths are expressed in terms of thousands of population, bushels per acre, wealth as so much per capita, expenses of operation in thousands of dollars of sales, etc.

Every ratio or coefficient has both a numerator and a denominator, the number or amount indicated by the ratio being in effect a comparison between the numerator and the

denominator. Ratios imply definite relations between the parts of which they are composed. If no such relation exists, or if the one established is "crude"—that is, general rather than specific—then the units of measurement are misleading.

To establish a coefficient, it is necessary (1) to secure the factor in the numerator, (2) to secure that in the denominator, and (3) to relate the one to the other. If any of these steps are not properly carried out, then the ratios or coefficients are faulty. And how frequently is the user of secondary data in doubt respecting not one but all of them!

A ratio or coefficient should be assignable to the conditions which make it possible. That is, the denominator should be capable of producing the condition named in the numerator. This is only another way of stating the thought of Bertillon when he says: "Always relate effects to the causes producing them."

One should not relate the number of deaths from spinal meningitis to the whole population, nor in this respect compare populations of entirely different age composition. Neither should one compare the number of industrial accidents for similar plants where the hazard or exposure, in terms either of man- or machine-hours, is widely different. Likewise, statistics of the number of farm accidents should not be related to the total number of farm employes, but only to the number employed in occupations producing the accidents. The mining industry is often classified as "dangerous," yet it is noticeably so only *when the accidents are related to the types of occupations in* which the hazard is exceptional.[1]

Loose thinking always results when effects are not related to the specific causes producing them. Long hours, poor ventilation and light in factory or mill are often assigned as the causes of occupational disease, yet it is not always clear how much of it ought not to be assigned to home life, intemperance, etc.—conditions only remotely associated with or en-

[1] For a more complete discussion of Units of Measurement, see Chapter IV, *infra*.

tirely dissociated from occupations *per se*. In each case, responsibility can be assigned only after investigation and after each effect is related to its specific cause.

It is not a sufficient justification for the violation of this principle to maintain that in economic life *effects* are rarely if ever to be attributed to single *causes*, and, therefore, that all effort to allocate the responsibility is useless. The statement is true but the inference does not follow. It serves, however, to call attention to the extra care which it is necessary to take in matters affecting economic and social conditions before conclusions are drawn from, and policies mapped out upon them. Again, the best that can be done here is to call attention to this important fact and leave the student, thus warned, to make application of it in each problem considered.

### 5. ARE THE DATA ACCURATE?

Accuracy is a relative term; it is impossible to secure absolute accuracy in measurements affecting social and business affairs. Some are more accurate than others, and so-called "accurate measurements" for one purpose may be grossly inaccurate for others.

The type of accuracy to which reference is made is not of the clerical type, although that is important. Computing devices which insure accuracy of this kind are now in common use, and it is seldom necessary, in using secondary data, to check numerical computations. Occasionally, however, errors of this type do occur.

"Sometimes they appear in the form of a disagreement of supposedly identical figures given in different numbers of the same journal, or of important inconsistencies in figures taken from the same table. Errors of this sort are, of course, sometimes due to misprints, which no care in publication can wholly eliminate. Sometimes, seeming inconsistencies are occasioned by the fact that preliminary figures are later subjected to decided revision. * * * But whatever their cause, the fact that significant discrepancies of various

types do occur indicates the need of careful examination of * * * data before they are utilized."[1]

The use of secondary statistical data is conditioned, among other things, by (1) the accuracy with which they are reported, (2) the accuracy with which they are determined, and (3) the accuracy with which they might be determined. Each of these different points of view requires brief consideration.[2]

The accuracy with which data are reported and collected depends upon (1) the type of informant, (2) the nature of the records kept, (3) the type of questions asked, and (4) the care used in answering them. If difficult and unfamiliar questions, or questions which in any way incite distrust or suspicion, are asked, answers are likely to be either incomplete, brief, noncommittal, general, or purposely evasive. Age, for instance, may be accurately known, but falsely reported. Wages may be known and yet incorrectly reported because of a suspicion as to the use to which the data will be put. Moreover, even in cases where there is no reason for data to be falsely reported, error may occur in transcribing and tabulating them.

On the other hand, data may be correctly reported but the report itself be inaccurate because the answer is wrongly determined. Much of the data, until recently, respecting causes of death fell under this head. No necessary difficulty is experienced in reporting,[3] but only in determining the precise cause, or in calling by the same name the same thing. The necessary corrective is, of course, the use of a standard classification of *causes of death*. Likewise, statistics of occupations suffer greatly from the lack of a standardized nomenclature. Identical occupations are called by different names; things

[1] Persons, Warren M., "Indices of Business Conditions," *The Review of Economic Statistics*, Cambridge, Mass., January, 1919, p. 6.

[2] For discussion of similar points respecting wage data, see Chapter V, "Types of Secondary Wage Data."

[3] See "Errors in Death Registration in the Industrial Population of Fall River, Massachusetts," *Monthly Review*, U. S. Bureau of Labor Statistics, Vol. 5, No. 1, July, 1917, pp. 2-8. This article slightly adapted is reprinted in the author's *Readings and Problems in Statistical Methods*, Macmillan & Company, New York, 1920, pp. 141-147.

which are equal to the same thing, in reality, are not equal to each other in name. As a basis for determining occupational risk, and for developing schemes of accident compensation or insurance, for instance, they are almost worthless. Fortunately, however, some progress toward uniformity of occupational naming is now being made. Here, as in the former case, the personal equation is important, but more often the real source of trouble lies, as in the instances cited, in the nature of the problem itself.

Statistics of "capital employed" in manufacturing industries, as reported by the United States Census Bureau, are faulty because of the inaccuracy with which they are determined. The definition of capital for statistical purposes offers the first difficulty. Authorities are not agreed as to what should be included as "capital." The reasons for including or excluding different categories vary and are of different force in different industries, or in the same industry under different conditions of management and forms of business organization. For census purposes, even, such a unit must of necessity be used with little more than a semblance of accuracy; and, of course, the statistics relating to it ought to be considered as estimates. The same thing applies to "value of products," "cost of materials," "expenses," etc. The difficulties are not necessarily due to errors in reporting (yet, undoubtedly, they are important), nor in the accuracy with which such facts *might be* determined, but rather with the accuracy with which they *are* determined under the conditions of collection.

If nothing more is desired than to indicate a trend, this may be done, in cases where complete accuracy of detail is wanting, provided errors are distributed uniformly about the average and tend to correct each other, and where sampling is representative. These conditions, however, so seldom obtain (never in the last instances cited) that data of these kinds must be used with great care for any use where accuracy is important. It is painful to see nice distinctions and weighty conclusions rest upon such questionable support!

On the other hand, secondary statistical data are frequently compiled where it is impossible to secure absolute accuracy, and where no pretense should be made that it is realized. The data at best are crude estimates. At present, for instance, no statistical machinery is available accurately to determine the amount of gold-producing ore in the United States; the horse-power of our water power resources; or the amount of standing timber in the United States.[1] Of course, there may be accurate as there may be inaccurate estimates, and it is always necessary to choose those which, all things considered, seem best to meet the requirements of the case. Moreover, they should be *used as estimates*. Essentially accurate conclusions may be drawn from rough estimates, if the basis upon which they are made is known, but even then, statistical skill and sound judgment are required.

Moreover, not all phenomena can be statistically measured. Numerical frequency may be of no real significance. For instance, the devotion of a people to a principle of right or justice can hardly be measured by the number of those who find no occasion to violate it. Neither can respect for law be determined by estimating or counting the number of people who remain out of jail. Conversely, disregard for law is not fully measured by the number of arrests and convictions. The number of those insane is not necessarily indicated by the commitments to insane asylums together with the occupants of such institutions. The sacredness with which marriage is regarded is not accurately reflected by the number of divorces granted; nor the number who are educated secured by totaling the students enrolled in institutions of collegiate and university rank. It is hopeless to expect statistical data alone to answer these questions.

[1] See the interesting report on "The Lumber Industry, Part I, Standing Timber," by *The United States Bureau of Corporations*, 1913, where methods of estimating the amount of standing timber in various districts and for various woods are described and criticized, pp. 7-10, 45 ff. This is reprinted in the author's *Readings and Problems in Statistical Methods*, Macmillan & Company, New York, 1920, pp. 91-110.

### 6. DO THE DATA REFER TO HOMOGENEOUS CONDITIONS?

Business and social relationships change—they are always in a state of flux. New policies, methods, and standards are always being introduced. New units of measurement, therefore, are needed to indicate the nature and extent of the changes, the old ones having lost their significance. The facts of yesterday may have little meaning for those of to-day. For instance, if, in a given market, "future" are supplanting "spot" transactions—and the level of prices has changed because of this fact—then prices of to-day cannot be compared with those of yesterday, when such methods of dealing were less common. Moreover, retail and wholesale prices cannot be directly compared. The conditions affecting them are different. Similarly, paper and gold prices cannot be compared until they are put upon the same basis.

Not only may statistical data be descriptive of non-homogeneous conditions (and this fact be not revealed), but they may also vary in composition at different times. Reporting, editing, tabulating, and analyzing may be of widely different degrees of excellence. Emphasis may have been differently placed; different definitions may have been insisted on; new units of measurement or modifications of old ones may have been employed; wider or narrower fields may have been covered; the proportional elements used to make up a total may have changed materially; etc. The presence of these and similar conditions makes comparisons over long periods difficult.

The desire for "comparability" often becomes the controlling factor in statistical computation, and serious omissions, strained interpretations, etc. (all important in the use of the data for a given time), countenanced in order to preserve it. For instance, the retention of the "capital" inquiry, in all its crudity, in the statistics of manufacture in the United States Census is largely out of consideration for the "value of comparisons." The omissions, until recently, of fifteen commodi-

ties formerly used in the computation of the index number of retail prices by the United States Bureau of Labor Statistics at least raises the question whether prices before 1907 can be compared with those since that date.[1] The various definitions of a "farm," of an "establishment," or of "manufacturing," as used by the United States Census Bureau at different times, make comparisons difficult over an extended period. Exports and imports, for instance, whether expressed in quantities or in values, must always be interpreted in terms of the units of measurement employed.[2] The student should always go be-

[1] The lack of comparability has been definitely asserted by a recent Commissioner of the Bureau of Labor Statistics. "Some Features of the Statistical Work of the Bureau of Labor Statistics." Royal Meeker, Commissioner, *Quarterly Publications of the American Statistical Association*, March, 1915, pp. 431-441.

[2] Most interesting discussions of the difficulties of making international comparisons of import and export statistics, and of the imperfections of our own import and export statistics, are contained in an article by Frank R. Rutter on "Statistics of Imports and Exports," in *The Quarterly Publications of the American Statistical Association*, March, 1916, pp. 16-35. Apropos the topic here under consideration, the following extracts are of interest:

By virtue of a law passed in 1893, the agent of a railroad company carrying goods to a foreign country by land was made punishable to the amount of $50 for failure to present a manifest to the collector of customs. "The effect of the change in law is reflected in the exports through Buffalo to Canada. From less than $500,000 in 1890 the figures jumped to over $4,000,000 in 1895." *Ibid.*, p. 20.

On the matter of units of measurement and classification, the following quotation is of interest: "The greatest need for the expansion of the classification is found in the case of exports. The most detailed classification of exports now covers less than 600 items, while in the imports for consumption there are about 3000 distinct items. The chief preventive of an increase in the number of items is the indefinite character of export declarations. So many articles are described merely by general terms that it is out of the question to separate articles frequently of much commercial importance.

"Defects in the present classification, aside from its incompleteness, are the incomparability of the import and export schedules and the failure to conform to current commercial terms. The latter defect is due to the preservation in the tariff of many terms now obsolete, and the necessity of having the statistical classes follow closely the tariff items." *Ibid.*, p. 26.

On the definition of "imports" the author says:

"What is generally understood by the term 'imports'? Legally, an article is imported when landed, whether for immediate consumption or for storage in bonded warehouses. From an economic point of view, however, bonded warehouses may well be regarded as foreign territory.

hind the printed figures and be sure of the units, their interpretation, and the weight assigned to the different factors in the composite groups before comparing them or using them as a basis for a conclusion.[1]

7. ARE THE DATA GERMANE TO THE PROBLEM BEING STUDIED?

He who has occasion to use secondary data, having inquired into the standing of the organization publishing them, and having satisfied himself as to the purpose for which they were issued, the nature of the data themselves, the units in which expressed, their accuracy, and the homogeneity of the conditions to which they apply, must then ask himself the fol-

The door of the bonded warehouse is really the economic frontier of the country.

"Since the United States is not a large reëxporting country, the difference between 'imports' and 'imports for consumption' is largely one of time. The instances in which goods are exported from warehouses are few as compared with the instances in which after the lapse of time goods are entered for consumption within the country.

"Perhaps the distinction is most clearly brought out by an illustration. While the last tariff was under discussion wool in large quantities was landed at our ports and stored in bonded warehouses until December 1, 1914, when it could be withdrawn without payment of duty. Was such wool really imported when it was landed or when it was removed from the warehouse?

"On the export side we have a clear distinction between domestic exports and foreign exports. On the import side imports for consumption are most nearly comparable with domestic exports, yet not fully comparable, since free goods are not generally warehoused and may be entered for consumption although intended for reëxportation. To be strictly accurate, dutiable imports for consumption should be compared with domestic exports and free imports with domestic and foreign exports combined." *Ibid.*, p. 28.

"Perhaps the most striking instance of the unfortunate result of our method of valuation is seen in the import prices of rubber. Notwithstanding the improvement of plantation rubber, Para rubber is still quoted at a slightly higher price. In Brazil, however, there is a heavy export duty, which constitutes an important element in the price. This duty is not included in our statistical valuation with the result that the value of India rubber imported from Brazil during the fiscal year 1914 averaged only 40 cents a pound, while the import value of that from Ceylon averaged 60 cents a pound." *Ibid.*, p. 30.

[1] Bowley, A. L., "The Improvement of Official Statistics" in the *Journal of the Royal Statistical Society*, September, 1908, Vol. 71, pp. 461-469, particularly. This article with slight adaptations is reprinted in the author's *Readings and Problems in Statistical Methods*, Macmillan & Company, New York, 1920, pp. 150-159.

lowing questions: *Are the facts really germane to my problem? Can they be used for the purpose which I have in mind?* These are significant questions. Upon the answers to them depend all subsequent steps in statistical procedure.

Many statistical data, which have only a general application to a particular problem, may, if used with discrimination, corroborate a thesis which they would not alone be sufficient to support. Contrariwise, they may be sufficient to throw suspicion upon, although they would not themselves disprove, it. *How data may be used, can never be known until their characteristics have been determined. They should never be used without this information.*

"The first thing to realise about official, and indeed all, statistics, is that their meaning is always technical and generally not precisely that which might at first sight be expected. * * * Statistics on any subject have generally a long history. In the beginning an organisation had to be initiated to collect records of those things connected with the subject which it was anticipated could be counted or measured. Experiment showed what facts could be ascertained and where the organisation was weak; criticism and analysis defined and interpreted the meaning of the totals and averages obtained, and showed their relation to the facts of which knowledge was desired. The organisation was gradually improved, new methods were devised for making good deficiencies, the meaning of the totals was modified and new definitions were necessary. When one has followed the process by studying successive reports or by reading a well-informed book or article on the subject the limitation and meaning of the totals can be appreciated; failing this, the best plan is first to think out for one's self what one would expect or wish to be included in a total (e.g. of the number of persons unemployed), then to read very critically word by word the heading, explanation and notes in the summary (always inserting some such phrase as 'recorded by' or 'reported to' or 'computed by' the department concerned), and then to get the larger report on which the abstract is based and study whatever information is there given about the method and purpose of the investigation. The critical faculty should be very alert when statistics are in question; the published heading may be pedantically and officially correct, but it will not contain such a statement as 'every word is used in a technical sense and has a special meaning only known to the officials who made the

compilation, the part that is not recorded is more important than that which is, where the facts are not known an estimate has been made by a method which cannot for departmental reasons be divulged, and the method of computation has been modified since *the last issue of the numbers,' yet part of all of this is sometimes* implied."[1]

In spite of the fact that statistics of some sort are to be found on almost every conceivable subject, those which are available may not suit the purpose in mind—if it is clearly *formulated—or they may apply to inappropriate times, places,* or conditions. It is then necessary to collect those which are suitable. Primary rather than secondary data must be secured. A discussion of the problems connected with such a task is the subject matter of the following chapter.

### References

Bowley, A. L., *Elements of Statistics,* 4th Ed., King, London, 1920, pp. 18-51.

Bowley, A. L., *Elementary Manual of Statistics,* MacDonald & Evans, London, 1915, Chapter VIII, pp. 64-70.

Bowley, A. L., *Official Statistics—What They Contain and How to Use Them,* Humphrey Milford, London, 1921, Introduction, pp. 7-13.

Chapin, F. Stuart, *Field Work and Social Research,* Century Co., New York, 1920, Chapter II, pp. 19-45.

King, W. I., *Elements of Statistical Method,* Macmillan & Co., New York, 1912, Chapters IV, V, VI, and VII, pp. 39-64.

Rugg, Harold O., *Statistical Methods Applied to Education,* Houghton Mifflin, New York, 1917, Chapter II, pp. 28-39.

[1] Bowley, Arthur L., *Official Statistics—What They Contain and How to Use Them,* Humphrey Milford, London, 1921, pp. 9-11.

# CHAPTER III

## COLLECTING AND EDITING PRIMARY STATISTICAL DATA

### I. Introduction

THE student or investigator who has occasion to collect pri-ıary statistical data must ask and answer the following uestions:

1) What is the precise problem upon which statistics are required?
2) Does the problem, as formulated, lend itself to statistical treatment?
3) What types of data are necessary for its analysis or solution?
4) Are they likely to be available in suitable form?
5) Are they likely to be adequate for the purpose in mind?
6) Will they have the required degree of accuracy, consistency, and comparability?
7) Can the data be made available within the time limit required: that is, will they have the required currency?
8) Are there likely to be any restrictions upon the use of data which will compromise the purpose which they are to serve?
9) What sanction is necessary, and what method of procedure, with the sanction available, must be followed in order to secure the desired facts?

Subsequent steps depend upon the answers supplied to them. 'hey constitute a sort of formal catechism to which one should e willing to subject himself before proceeding further.

### [. Preliminary Conditions to the Collection of Primary Data

The problems involved in satisfactorily answering each of ıe above questions require separate consideration.

### 1. WHAT IS THE PRECISE PROBLEM UPON WHICH STATISTICS ARE REQUIRED?

The idea of a *problem* suggests a difficulty—some thing or an aspect of a thing which is unsettled or not understood. Before it can be stated, it must be *clear* that there is a *problem*. Its precise nature will take form as its different aspects are contemplated—that is, as they are mapped out and delimited. *Such contemplation is thinking.* To *think* on a problem is to survey the facts about it; to define and classify them, and to see them in their proper relations. Not until it is known what facts are to be considered and in what way can a problem be stated; and once it is *stated clearly*, its solution is greatly expedited.

To think about a problem and to state it require *knowledge* concerning it. This has to be acquired: it does not simply "come." It is sheer waste of time to begin collecting data on a problem until it is defined. It must be seen in relation to other problems. What these relations are can be known only by *thought* about them.

The first and most essential step concerned with the collection of statistics with respect to any problem, therefore, is to define and state the problem itself.

But all problems do not lend themselves to statistical study. Some do; others do not. Moreover, when the statistical approach is used, it is not always used in the same way. Neither does it involve the use of the same methods. Accordingly, the second question which must be asked before data are collected is:

### 2. DOES THE PROBLEM, AS FORMULATED, LEND ITSELF TO STATISTICAL TREATMENT?

Statistical studies are necessarily quantitative; statistical facts are always numerical. Moreover, the frequencies or attributes of things have to be sufficiently distinct so as to make it possible to enumerate them. It is possible, for in-

stance, to study statistically the sex and age characteristics of insane inmates in hospitals; it is not possible by this method to determine the fact of insanity. It is also possible to measure the distribution of the wealth and income of a people, but it is not possible by statistical means to determine what distribution is socially or politically desirable. Again, if bankruptcy is considered equivalent to business failure, then the number of such failures, by types of business, age of business, location, capital investments, liabilities, etc., may be statistically determined. The parts which dishonesty, moral cowardice, speculation, etc., play as contributing factors to such failures, however, cannot be directly measured in this manner. Why? Because they cannot be numerically stated, and if they could, their significance would not be indicated by quantitative preponderance.

A problem to be susceptible of statistical study should have characteristics which are quantitatively measurable. Moreover, they should be capable of being distinguished with respect to time, to place, or to degree. Such conditions hold, for instance, for prices, wages, deaths; they do not obtain, for instance, for integrity, honesty, loyalty.

If it is decided that a problem may be studied statistically, and for this purpose it is necessary to collect primary data, the next question which the investigator must ask himself is:

### 3. WHAT TYPES OF DATA ARE NECESSARY FOR ITS ANALYSIS OR SOLUTION?

The answer to this question depends upon (1) whether data are needed to supplement, corroborate, or disprove those already available upon the subject, or (2) whether an entirely new and different set of facts, expressed or measured in new units, is necessary. If the former condition obtains, then the data collected must have the same characteristics as those with which they are to be compared, or which they are to supplement. They may apply to different times and places provided they exhibit themselves in the same way. Indeed, the

types of data already available on a problem determine the nature *of those which are collected to supplement them.* To duplicate what has been done is justifiable only when it is felt that existing data are incomplete, unrepresentative, or in some other respects inadequate or unsuited to the uses to which one desires to put them. The aim should be to supplement, to carry further the type of analysis which has already been made—to make the data already available function. Too frequently, statistical studies are uncorrelated with those already existing. They cover old ground, and contribute little or nothing to an understanding of the problems with which they have to do, largely because they do not constitute a necessary part of a comprehensive program, nor dovetail with the studies which have already been made. They begin and end as isolated, unrelated efforts.

If, on the other hand, data are to be collected but not to supplement those already existing, then choice is free, but within clearly defined limits. The first question which is presented is:

4. ARE THEY LIKELY TO BE AVAILABLE IN SUITABLE FORM?

Data which exist may not be *available.* They may be (1) confidential, (2) expressed in units unsuited to a particular use, or (3) scattered over so long a period or over so wide a territory that the expense involved in their collection is prohibitive. *Another question is:*

5. ARE THEY LIKELY TO BE ADEQUATE FOR THE PURPOSE IN MIND?

A satisfactory answer to this query can be made only if the purpose *is* known, and if means are available for knowing the probable nature of the data. It is taken for granted that the first condition is fulfilled; the latter may be satisfied by sampling the data, or by consulting with those who possess them.

### 6. WILL THEY HAVE THE REQUIRED DEGREE OF ACCURACY, CONSISTENCY, AND COMPARABILITY?

The types of records from which they are to be drawn, the honesty of the informants, the care with which they are transferred from the original records, and the manner in which the information is solicited all have significance in this respect.

It is necessary to know the types of informants to whom appeal must be made. If they are ignorant, inclined to depreciate the significance of the problem under study, or to oppose its continuance; if they are inclined to look upon everything as inconsequential and useless, little weight can be attached to the answers given. Likewise, the time, money, and organization available should be considered. Data may exist, informants be ever so willing to supply them, and yet the necessary facts be unavailable because of lack of funds or of time in which to secure them. Few people, not accustomed to planning statistical work, clearly realize the time, energy, and expense involved in a thorough statistical investigation.

Further questions must also be asked and answered before the task of collection is begun. One which is important is as follows:

### 7. CAN THE DATA BE MADE AVAILABLE WITHIN THE TIME LIMIT REQUIRED: THAT IS, WILL THEY HAVE THE REQUIRED CURRENCY?

On some problems, data to be significant must be current. This is true when they are needed to determine present rather than to reflect past conditions. On the other hand, for the solution of certain problems, current data are of less value than those which refer to the past. If, for instance, the *normal* relation between sales and expenses is to be determined, then current data are inadequate. Those which have to do with a *normal* period in the past are necessary.

When matters of current business or of social interest are pressing for solution, statistical data referring to the past

are heavily discounted. The fact is, however, that the policies of to-day grow out of those of yesterday, and into those of to-morrow. The past is the present viewed in retrospect; the future, the present viewed in anticipation. The desire always to be "up to date" amounts in some instances almost to a mania. Sober thought of the past is often stabilizing, serving as it does to give a proper perspective to the present.

### 8. ARE THERE LIKELY TO BE ANY RESTRICTIONS UPON THE USE OF DATA WHICH WILL COMPROMISE THE PURPOSE WHICH THEY ARE TO SERVE?

Restrictions may take a variety of forms. For instance, certain data (1) can be published only as totals, or the instances only in groups; (2) cannot be published at all; (3) cannot be distributed except to a select few; or (4) if published at all must be given general distribution.

### 9. WHAT SANCTION IS NECESSARY, AND WHAT METHOD OF PROCEDURE, WITH THE SANCTION AVAILABLE, MUST BE FOLLOWED IN ORDER TO SECURE THE DESIRED FACTS?

Most public agents are possessed of mandatory power: that is, they may compel answers to be made to questions asked. Private individuals do not usually have the same sanction and its absence in most instances is a handicap. It is, however, sometimes possible for investigators, through contact with informants, and by cultivating their good-will, to develop in them a feeling of obligation to report, which more than compensates for any lack of mandatory power. So far as public statistical organizations are concerned, conspicuous instances, where a feeling of obligation to supply information has been well developed, are the cases of price reporting to the United States Bureau of Labor Statistics, and the reporting by unions of the conditions of employment to the Bureau of Labor Statistics in Massachusetts and in New York.

By cultivating the good-will of informants, these bureaus have been able to enlist their support, and to secure excellent data with little actual inconvenience and cost. Various ways are open for securing their interest and good-will. One method is to guarantee that confidence will not be abused, that the study is scientifically undertaken and without the idea of personal gain or aggrandizement. Sometimes it is accomplished through assurances being given that the request for statistics extends to a whole class rather than to a selected number of a class, and that when the returns are compiled they will be supplied gratuitously to all those who have contributed to their collection. Sometimes an effective method is to appeal to feelings of state or local pride, or to class conscious sentiments.

Another way of gaining the confidence of informants is to study their interests and to cultivate their good-will by correspondence. This method is being used effectively in Massachusetts, where bureau officials are careful to indicate by semi-personal letters the value to informants and to the public generally of data to be collected, and the importance of answering specifically and promptly the inquiries made. Even where mandatory power exists, it is not an uncommon practice for statistical bureaus requesting information, while quoting the terms of the law under which the collection is made, to make the idea of co-operation their chief appeal. A display of force or the use of threats should be used with discrimination, inasmuch as it may tend to incite a spirit of distrust and opposition rather than of co-operation.

Private individuals, as contrasted with regularly constituted authorities, are usually handicapped in the collection of data by lack of sufficient sanction. The limitations under which they operate should be clearly kept in mind in order to guard against a too sanguine belief that they will always secure the information desired. Too great confidence as to the outcome of a given undertaking generally characterizes the efforts of the inexperienced.

## III. The Collection Process

### 1. PURPOSE AND PLAN

The process of collecting primary statistical data depends upon the purpose in mind and the plan outlined to realize it. There can be nothing hazy, confused, or indefinite about them if satisfactory results are to be secured. The problem should be clearly thought through and the plan be made complete from beginning to end. Only by so doing is it possible to provide in advance for the contingencies which are sure to arise. Both require thought and care. Rarely, if ever, can statistical studies be rushed. Progress is made slowly. An adequate foundation respecting both purpose and plan is essential. They are so important that much of Chapters IV and V is devoted to a discussion of them for typical problems.

### 2. THE COLLECTION PROCESS DESCRIPTIVELY CONSIDERED

The ways in which data are collected vary with the nature of the problem, and the organization which undertakes the task. No two problems require exactly the same methods. Each has its peculiar requirements. In every case there is a *best* method, and it is part of the task of the organization to determine what it is under the conditions obtaining.

Statistics, like other information which is desired, must be secured by *some one*, in *some way*, according to *some method*, and from *some source*. The one securing it may be the agent —private individual or organization—or his representative. The way in which it is solicited may be by interview, by personal letter, by questionnaire or schedule, or by all of these means. The method of securing it may involve a count or an estimate, and the source of both may be found in personal opinion, or in records.

The simplest situation in which data are collected is probably that in which an organization or business merely summarizes or assembles information about its own activities. The collection may involve data currently kept in systematic

form in its own records, or it may pertain to facts not a matter of record but of estimate or opinion. Examples of the first type are sales, expenses, profits, output, loans, capital, assets, number of employes, etc. Illustrations of the second kind are estimates or opinions of salesmen respecting sales prospects for the coming season or year, general business conditions, influence of competitors, etc. In problems relating to matters of record, adjustments in the form of accounts, units, etc., are necessary where the methods are not standardized in the different departments. In all cases of this sort, however, it is assumed, after the plan is thoroughly worked out, that so far as the collecting or assembling of the facts is concerned, the task is largely one of transcribing in suitable form the data available. Motives for withholding part of the facts, of inaccurately stating those supplied, or of attempting to defeat the project, are not generally present. Unity of management tends to guarantee against failure in these respects.

Moreover, personal bias, the desire to make a case, or reliance on incomplete data do not normally obtain under such conditions. Of course, data assembled in this way are not always adequate for the purposes in mind. They may be incomplete, and inaccurate for other reasons than those suggested, more particularly if the assembling is done under the direction of some one untrained for such work. But collection under such circumstances does not present the problems which confront the statistician from the outside who attempts a similar task, and who has no other sanction than that of an impersonal government or his own good intentions, and who too frequently has not the tact to enlist the sympathy and cooperation of those upon whom he must depend for success.

It is, of course, true that most smaller business houses do not understand the uses to which their data can be put, and consequently do not have satisfactory statistical records. Moreover, those who appreciate their possible significance may have considerable reservation about giving over to a separate

department the responsibility of informing others of the weak places in their organizations. "Statistics" are often in ill repute because they are considered either in themselves infallible or fallible—depending on whether they show the right or wrong thing—or because they are used unscientifically. There is almost as much science in the way statistics are collected as there is in their subsequent use, but this truth is rarely appreciated by the inexperienced.

More difficult situations in collecting data are encountered when information, although a matter of record, is desired about business, trade, or social phenomena by some one from the outside. The nature of the records is frequently unknown, and direct access to them impossible. If they are furnished in the original, adjustments, corrections, and interpretations have to be made after they are received. If their contents are transcribed by an informant, they have all of the limitations possessed by the originals; if by the agent soliciting the information, they must be taken in the form in which they are found or adjusted in keeping with his idea of appropriateness. For an agent to tamper with original records is a dangerous procedure. The meaning of the facts may be confused; they may be wrongly interpreted and combined in ways in which they were never intended.

To permit informants to transcribe their records is expeditious, but the liberty may be construed as license. In some instances, requests for information may be ignored, or answers given which are evasive or susceptible of different interpretations. Unless there is some check upon the information supplied, this method of securing data is inadvisable for general use.

Where questionnaires are used and informants are required to fill them out, the answers to questions may be incorrect because the questions (1) are misunderstood, (2) call for information about which little or nothing is known, and (3) use units of measurement which are unfamiliar. Long explanations cannot conveniently be made upon questionnaires, and

if they are supplied, no attention may be given to them. Only when informants feel obliged to answer questions, or where answers may be thoroughly checked can complete reliance be placed in information supplied by schedules which informants themselves have filled out. In the investigation into "Wages and Regularity of Employment in the Cloak, Suit, and Skirt Industry, etc., in New York," the information, supplied upon 1429 schedules filled out by the workers and gathered by the shop chairmen, was found to be "so full of errors that they were discarded as entirely unreliable."[1]

So much for a consideration of the problems of securing information, which is made a matter of record, when it is assembled by those within an industrial or other business, or when collected by those from the outside.

On the other hand, information is frequently desired about conditions which are changing. Each time it is wanted, the phenomena with which it is concerned must be separately observed. The following are illustrations: inventory stocks on hand, the population of cities, people passing a given corner, daily receipts of cattle at Chicago, etc. To secure such aggregates, the instances must be counted, the act being repeated each time data respecting them are desired. Records of past events may have a certain significance as tests of the accuracy of a given enumeration, but of and by themselves, they do not supply the information that is desired. A photograph, as it were, must be taken of the phenomena at the time in question and for the area or conditions involved.

The nature of the problems involved in *a count* will be evident from a consideration of a typical case. An example in which counting is required, is the enumeration of the population of the United States. The excess of births over deaths, together with the surplus of immigration over emigration, are the sources making for an increase of population. Reasonably accurate statistics of births and deaths are restricted in the

[1] *Bulletin of the U. S. Bureau of Labor Statistics*, Whole Number 147, p. 14, Washington, D. C., June, 1914.

United States to the so-called registration area. Statistics of immigration and emigration are reasonably accurate for the country as a whole. Statistics of distribution of immigrants, more accurate than possibly the state to which they *declare* they are bound, or of the origin of the emigrant, more definite than his last place of residence, are not available. Little or no record is kept of migratory movements of population within the country. The result is that for statistics of population, reliance must be placed in the decennial census made by the United States Bureau of the Census.

The actual enumeration of the population of 110,000,000 people in a district as large as the United States is a gigantic undertaking. Even if the tendencies for districts to exaggerate their figures and for enumerators to pad their lists in order to increase their remuneration are ignored, the difficulties are almost insuperable. Coupled with these conditions, and serving the political purpose which a census does, little value so far as absolute or even near accuracy is concerned can be attached to it as an actual enumeration or count. With the reasons for this state of affairs, attributable as it is to the method of appointing enumerators, to the inherent size of the task, to the divided duties of the enumerators between a population census proper and an agricultural and occupational survey, to the political purpose which it serves, etc., we are not here particularly concerned. Our chief interest is in the method rather than in the accuracy of the data collected. Questions involving the determination of legal residence, the treatment of floating population, of people in transit from place to place, etc., are involved in the process of counting.

In the case of a population census, partial checks on the accuracy of the count are found in the preceding censuses, in the records of deaths, births, immigration, emigration, and in the fact that normally the distribution of age and sex classes is essentially uniform from period to period (this relationship is somewhat disturbed in the United States by the

influx and egress of mature male immigrants). These checks, however, valuable as they are to keep in bounds of reasonable inaccuracy the results of the canvass, do not, even under the best of conditions, lessen the inherent difficulties of counting large aggregates even with approximate accuracy. The frequency of contested elections, in cases where crookedness is admittedly absent, furnishes another evidence of the difficulties in correctly counting large aggregates.

Not only may *actual* instances be recorded and *actual* cases be counted, but the probable frequency of their occurrence or appearance may be estimated. Estimates may be made on the basis of what has occurred in the past, or on what is likely to occur in the future. They may be made on the basis of *direct material,* as when expectancy of death (life tables) is based upon the number and conditions of deaths. They may also be made from *allied material,* as when call-loan rates of interest are estimated on the basis of bank reserves, the net interior movement of money upon the size of crops, the trend of business on the combined factors making for business distrust or confidence, the probable price of corn upon the price of wheat, etc. Indeed, in the business world most dealings are hazarded upon the ability to foretell the most probable results from a given set of conditions. Market prices of cereals are, in large part, a reflex of the likely condition of croppage during the subsequent six or twelve months balanced over against the likely conditions of demand; prices of securities are based upon an estimated earning capacity of the properties floating them; increases of investment are hazarded upon a continuance of favorable trade conditions, or the favorable disposition of the legislature, etc.

Much of the statistical data regularly compiled on the agricultural outlook; on the depletion or conservation of resources; upon national wealth and its distribution; upon the benevolence or malevolence of a given state policy toward business and industry, or the likely consequences of the adoption of a régime of Socialism or government ownership; upon the dele-

terious effects of a given work policy or condition upon the laborer, etc., are estimates. Some of the data are sufficiently accurate for all practical purposes, are compiled under conditions which tend to give them value—since absolute accuracy is unnecessary—and may serve as bases upon which to formulate a policy or launch a program. Such, undoubtedly, is true respecting the data issued by the Department of Agriculture at Washington on the condition of crops, on the acreage of cereals, etc. Absolute accuracy is not required, and the amount of error, tending as it does widely to distribute itself and to remain essentially the same from period to period, is not a seriously disturbing factor.

On the other hand, estimates made respecting conditions which constantly change, and upon which adequate data as guides do not exist, or which in themselves are impossible of determination, have serious limitations. Too free use should not be made of them in shaping governmental or business policies and in questioning social and economic institutions. The estimated amount of arable land in the United States is materially increased by the completion of irrigation projects and the perfection of dry-farming methods. Power sites available are increased in number and value by the perfection of high-power transmission apparatus, and the available supply of precious metals, by the discovery and use of the cyanide process for separating gold from crude ores. The estimated fuel supply takes on new significance in the light of recent discoveries respecting the use of oils and the perfection of internal combustion engines. The partial displacement of the steam by the gasoline engine puts in a new light the consequences which are sometimes associated with an estimated rapidly diminishing fuel supply.

We are, however, not concerned at present with the consequences of a condition, the facts about which are arrived at largely, if not wholly, through estimates, but rather with this method of numerically describing such condition or tendency. Attention is simply called to the fact that a very large pro-

portion of statistical data currently collected by government and private statistical bureaus is nothing but estimates. They may be good, bad or indifferent; but this does not now concern us. They should, however, be used as estimates, and the limitations of the methods under which they are collected be fully understood.

Whether recorded information is used, or counts or estimates made, depends upon the problem in question, the nature of the data needed, and the form in which they occur. In these respects, each problem will be differently handled. Descriptively, the methods differ.

### 3. THE COLLECTION PROCESS FUNCTIONALLY CONSIDERED

In collecting data—irrespective of the type which is desired and the precise methods which are used to secure them—there are, however, conditions which have universal application. There is a fundamental technique of use, usable in all cases. It is a *function* of all methods, although it is descriptively different in each. It has to do with (1) the source of material, and (2) with the manner in which it is secured.

#### *(1) Who are to be Canvassed?*

As soon as the purpose of a statistical study is stated, the following question immediately arises: From whom and in what way shall the data concerning it be secured? The first problem, stated in another way, is: Who shall be canvassed? A preliminary answer to this question can be given by a hurried survey of the problem and an inspection of the sources available. A complete and definite answer is possible only after a list of the possible sources of information has been made and the types of the informants, together with the character of the material which they possess, determined by careful study. To illustrate: If the problem is to fix a reasonable minimum wage for gainfully employed women, inquiry about the wage scale in use must be directed to those who clearly

fall within the group affected. If the wage is to apply to a single industry, then obviously there is a double restriction imposed.

Having determined the industry and the persons affected, however, the question remains: From whom shall information be secured? If the prevailing wage-rate is secured from employers alone, objections may be raised that the returns are inaccurate; that all cases are not included; that the data apply to unrepresentative seasons; that the money value of perquisites granted are included in the wages reported; that because of the stability of employment and the security of tenure, these factors are capitalized and included as a part of the wage or counted as equivalent to monetary compensation, etc. If the same facts are secured from the workers alone, the contention may be made that records are not kept and, therefore, that the data submitted are at best estimates; that no cognizance is taken of other things than money wages, and that there are evidences in the data submitted of a desire to make a case. Neither source may be depended upon absolutely. In case there are irreconcilable differences in the reports or testimony submitted, reported figures in the absence of the actual facts will have to be taken. If any of the above considerations obtain, they, of course, may be given weight in the determination of actual conditions. A single source is not always adequate; it is frequently necessary and desirable to use various sources in order to get the facts and to see them in their correct light.

Again, if the subject of study is budgets of workingmen's families, such questions as the following will have to be answered: Who are workingmen? Who shall be included and who excluded in a particular study? What national, racial, customary trade, occupational, and wage boundaries shall be set up? How many budgets can be secured? How many are needed and what periods must they cover in order accurately to characterize the situation? How wide must the survey be to be typical of the group or class? Such ques-

tions cannot be answered offhand. The way in which they are asked and the use which is made of the answers received require careful consideration and the use of keen judgment and sound statistical sense.

In order to measure the effects of a law which requires all employers of five or more persons to report industrial accidents to a central authority, and to make conditions of labor safe by the adoption of adequate safety devices, it is necessary to know who are affected by its provisions. Failure to comply with a law cannot be made punishable when the supplying of blanks for reporting accidents and recording the installation of safety devices, for instance, is made a condition of the law's operation, and this the administrative board has failed to do. In the administration of such laws, one of the most difficult problems is the preparation and current correction of lists of those to whom the law applies. A statistical statement of the results accomplished or of the conditions obtaining in industry is impossible without a determination of those who are affected.

Not infrequently, conditions of time, money, and organization require that sources of information be omitted or that typical facts alone be presented. The problem then becomes one of sampling. What shall be used and what omitted? An index number of prices may be materially affected by the omission, or by the too frequent use of a given commodity or of certain types of commodities. The reasonableness of a court decision, or of an administrative ruling as to what constitutes a "fair return" upon railroad property, may hinge upon the inclusion or exclusion of certain representative railroads. The omission of an important sale, under the sales method of real estate valuation, may affect the value given to real estate in a given district. In the determination of a unit-value for urban land, how much importance shall be assigned to corner influences, to frontage, and to relative position? Small deviations from the standard usually employed may make a large difference in the value assigned. The area included may be too

large, conditions may not be homogeneous, and the resulting unit-value not be typical. The problem is essentially one of judging the conditions to be included, and of determining the weight to be assigned to each controlling factor in order that the sample may be truly representative.

Who shall be canvassed, and what conditions shall be included, depend in large part upon whether samples will suffice, or whether all data are necessary for an adequate picture. If it is decided to employ samples, care should be used (1) to distribute them over as many categories as are represented in the complete data, (2) to include them in proper proportions, and (3) to guard against an undue emphasis being given to any particular quality or feature peculiar to a given type or class.

Comparatively few workingmen's budgets, if accurately kept and reported, will serve to give a correct picture of the cost of living.[1] It is unnecessary to include all individuals of the class considered. The Bureau of Statistics in Massachusetts maintains that the returns from representative manufacturing establishments are superior to those which would be secured if returns from all establishments were included. What is desired, of course, is not a record of capital employed, wages paid, etc., for *all* establishments, but only for *representative* ones. On the other hand, in the collection of statistics of trade union membership and the amount of unemployment, it is necessary to get totals for all unions. No reasons exist for the use of samples—the statistics are meant to be inclusive. If they are not, the only alternative is an estimate upon the basis of the incomplete returns.

*Functionally*, such questions as those just presented apply to all problems upon which data are to be collected. The precise methods used to secure the facts vary. *Descriptively, they are different; functionally, they are the same.*

[1] For an interesting discussion of sampling, see *Livelihood and Poverty*, by Bowley, A. L., and Burnett-Hurst, A. R., London, 1915, Chapter VI, pp. 174-185.

### (2) *The Ways in which Primary Data May be Secured*

#### a. Personal Interviews

In business and in some social surveys it is a common practice to secure information by interviews. Personal contact is established and the required data are solicited first-hand. Whenever this method is employed its success depends, among other things, upon

(1) the sanction possessed by the person making the interview.
(2) the personal qualities of the interviewer—his tact, diplomacy, courage, and intellectual curiosity.
(3) the degree to which he understands
  (a) the problem upon which he desires information.
  (b) the psychological and instinctive reactions of those whom he interviews.
(4) the accuracy with which he
  (a) interprets the information supplied.
  (b) records or remembers the facts submitted.
(5) the form of record upon which the answers are put.

#### b. The Use of Form or of Personal Letters

Success or failure will attend the efforts of those seeking information by the use of form or personal letters in proportion as they

(1) inquire of those who have the desired facts.
(2) are definite and precise in stating what is wanted.
(3) ask for data which are a matter of record rather than of opinion.
(4) formulate their inquiries in such a way that the units in which the data are measured are
  (a) the same as those which are currently used.
  (b) not overlapping.
  (c) simple rather than being composite or expressed as ratios.
(5) are able to overcome the natural indifference and reluctance to give information which is
  (a) confidential.
  (b) difficult or costly to assemble.
  (c) of use to active or potential competitors.

(6) are able to reciprocate in some way or make their finding accessible to all.

### c. The Form, Use, and Editing of Questionnaires or Schedules

Questionnaires may be used either with or without personal interviews. Used in either way, they constitute (1) the list of questions which are to be asked, and (2) the form upon which the answers are to be recorded. If they are personally distributed, and their filling in is done by the agent himself, or under his direction, the purpose and nature of the inquiry to which they relate, as well as the terms used, may be explained, doubtful points cleared up, and corroborative questions asked. If, on the other hand, they are distributed by mail and filled out without assistance, then they themselves must carry conviction, be self-explanatory, consistent, and persuasive. Personal appeal for information, best made by human contact, is then made through the printed rather than the spoken word. Of course, objections to giving information, indifference and apathy on the part of those having the desired facts may be dispelled by personal contact in advance of the distribution of the schedules. But this is rarely possible. Those who have the information are generally too numerous and too widely dispersed to be influenced in this way. Complete reliance must be placed in the questionnaire itself. Since this is necessary, the only way in which the end may be accomplished is to make the questionnaire adequate for the purpose. Accordingly it is well to observe the following principles of schedule making:

(1) Assurances should be given that the inquiries are made according to the provisions of law, or if voluntarily undertaken, with the hope of throwing light on some particular problem. Reasons for making the inquiries, and for making them of the particular informants, should either be stated or be clear by inference. Informants generally demand assurance that the

law requires answers to be made, or that the purpose sought to be accomplished has some really vital end.

(2) Questionnaires should be accompanied by stamped envelopes for return.

(3) They should be as brief as is consistent with the purposes which they are to serve, and the questions asked should unmistakably be addressed to the problem. So far as possible, the significance of each question should be evident from its context.

(4) Units of measurement should be clearly indicated, be accurately defined, and conform to common usage. Definitions and explanations should appear in the body rather than at the beginning or the end of schedules.

(5) Rulings and columnar arrangement should be simple and definite so as to guard against the misplacing of items. If spaces or columns are not to be used, this fact should be clearly indicated.

(6) The page should not be crowded, ample space being provided for all answers. Related questions should be grouped together.

(7) Opportunities or occasions for making false or inaccurate answers should be guarded against by having the questions, so far as is possible, corroboratory.

(8) As a rule, the making of arithmetical calculations as totals, percentages, etc., should be reserved for the statistical organization, and not intrusted to or imposed upon informants.

(9) Questions should be simple and unmistakable as to meaning, should not allow of evasive answers or of double interpretation, should not be unduly inquisitorial, should be arranged logically and in the order most convenient for the informant, should not involve duplications, should be capable of being answered by "yes" or "no," by number or amount, and should always be courteous and diplomatic in tone.

The sending out, returning, and editing of questionnaires raise some interesting problems which call for brief consideration. As a rule, all questionnaires should be sent out at the

same time. If this is done, it will tend to allay a suspicion which may arise in case one of a group receives his copy in advance of others. He may feel that he is being singled out for special inquiry. Moreover, the simple expedient of sending out questionnaires simultaneously tends to guarantee against their being late in returning, and interfering with the process of tabulation and analysis. If returns come straggling in, it is often difficult to know when to "close," and what to do with late returns. Repeated requests may be made for information, but the amount of pressure which can be applied in case of a failure to report, as well as the success which will attend such efforts, will depend upon (1) the importance assigned to a given return or to additional information, (2) the mandatory power possessed by the inquirer, (3) the degree of co-operation which obtains between the informant and the person or organization seeking the information, and (4) the period available for delays, and the position arrived at in the process of tabulation and analysis.

When schedules are returned, whether this is done by informants, or by representatives of the collecting agent, a certain amount of checking, editing, and revising is necessary before they can be accepted and tabulation begun. If agents of the collecting unit send them in, they will be uniform in most details, and occasions for correspondence and personal interview regarding the meaning of certain entries obviated. The services of agents in such cases will have been used in making the entries rather than in correcting and adjusting them after the schedules are received.

Evident errors due to omissions, additions, false entry, confusion of items, etc., can be readily corrected. Undue tampering with the facts, however, is dangerous. Alterations should be made only in cases of unmistakable error. It is an easy matter materially to change the meaning and to distort the truth of answers by the interchange or erroneous correction of a few items. The will to deceive may not be present at all, and yet the same results follow as if it were. If questions

have been uniformly misunderstood, the basis for change is certain. If, however, the relations between items are made to agree with what in the editor's opinion "ought" to be the case, then the data are used merely to support individual opinion.

The degree to which omissions may be allowed or error countenanced is also of importance. If entries tend unmistakably to confirm an ascertained fact, and the samples are representative, then the omission even of a number of questionnaires may be tolerated. If, however, the evidence is uncertain or conflicting, the trend or the relations being indefinite, then the omission of an item in a comparatively few cases may be a serious matter. It may be that these are the very items which are needed to decide the case in point. No rule of tolerance can be formulated which will cover all such cases. If the range for discrimination is wide, or discretion given too wide a latitude, final results may be determined quite as much by the judgment of the editing official as by the data themselves.

Many of the same considerations apply in the case of error. If errors tend to correct each other, a considerable degree of inaccuracy may be allowed. If, however, they tend to become cumulative, then their presence is of serious consequence and every effort should be made to remove them.

These different aspects of editing may be illustrated by considering the uses of the "sales method" of determining real estate values. All biased errors must first be removed. These are interpreted to include, among other things, sales involving nominal considerations; transfers between relatives; and land contracts or other conditions which in any way cloud the titles. Only sales between ready and willing buyers, and ready and willing sellers, and accompanied by full warranty deeds, are held to be valid for this use. By eliminating "doubtful" sales, however, the number actually available as a basis for deciding what the value is in a particular district may be inadequate. If this occurs, then shall sales made

between relatives, when the values represented by them essentially agree with the findings when they are omitted, be included? Provided the value thus determined is warranted, to use them would tend to confirm the value arrived at on the basis of other sales. If it is not warranted, then their inclusion supports a conclusion which in and of itself is incorrect, and weight would need to be given to the conditions under which the sales were made. Their inclusion, on the other hand, may materially change the values assigned to a given district, and yet, from the evidence available, it may be clear that they represent true value. The only consideration against their use is the relations of the grantees and grantors —relations which normally would make it inadvisable to use them in order to determine land values.

Moreover, how many sales are necessary to establish a unit value? With twenty sales, the unit value might be $100 per front foot; with twenty-five sales, $105, and with eighteen sales, $95. How many sales should be included?

Such considerations as these are involved in every statistical problem and in the collection and use of statistical data, no matter whether they apply to land valuation, price determination, studies of wages, cost of living, or what not. To edit primary data requires sound judgment and keen discrimination.

## IV. Conclusion

This chapter has had to do with the collection of primary data and with their preparation for use. The discussion is intended primarily as a manual of instruction rather than as an encyclopedic treatment. If the points of view developed are kept constantly in mind, and there is real desire to profit by them, subsequent steps will be easier and the reader will have the assurance that he is employing in a scientific manner a delicate, though frequently abused, method of induction —statistical methods.

The personal element stands out as an important factor in

all that has been said. Statistics do not answer questions or support conclusions independently of those who manipulate them. Judgment, candor, and integrity are necessary at every step. One must know the field in which he is working, its statistical possibilities, and what has been done. He must also realize the difficulties under which data are collected, the precise manner in which they are to be used, the sources and possibilities of error and bias, etc., and the ways of detecting and eliminating them. In a word, he must understand what is involved in the preparation of an intellectual tool, and then in the light of his knowledge use it intelligently for the purpose in mind. If it is faulty, he should know and acknowledge it. If it is well fitted for his purpose, that fact should be evident in the uses which are made of it. To be a good statistician one has to be more than a technician, but technique cannot be ignored.

## References

Bailey, W. B., and Cummings, John, *Statistics*, A. C. McClurg, Chicago, 1917, Chapter III, pp. 8-16; Chapter IV, pp. 17-25.

Bowley, A. L., *Elements of Statistics*, 4th Ed., King, London, 1920, Chapter II, pp. 14-17.

Chapin, F. Stuart, *Field Work and Social Research*, Century Company, New York, 1920, Chapter VII, pp. 148-191.

Rugg, Harold O., *Statistical Methods Applied to Education*, Houghton Mifflin, New York, 1917, Chapter II, pp. 39-56.

## CHAPTER IV

## UNITS OF MEASUREMENT, OF ANALYSIS, AND OF PRESENTATION IN STATISTICAL STUDIES

PASSING from the more general statement of the methods of collecting statistical data, and of the principles involved in the collection process, the significance of such expressions as *units of measurement, of analysis, and of presentation* will be clearer if they are discussed separately in connection with concrete problems. This is done in this chapter.

### I. THE MEANING OF STATISTICAL UNITS OF MEASUREMENT

The statistical approach to a subject is numerical. Things, attributes, and conditions are counted, totaled, divided, subdivided, and analyzed. It is concerned not with single instances or with rare occurrences, but with aggregates.[1] The statistical process requires both analysis and synthesis, numerical preponderance being the chief basis for conclusions based upon such aggregates.

Statistical frequencies or amounts relate to units of measurement which are characteristic of the things or conditions studied. It is not 1000 as an abstract unit, but 1000 farms, industrial establishments, loans, and mortgages, which are considered. *Abstract* numbers or frequencies, on the other hand, may be combined, separated, and divided in an infinite number of ways because they are homogeneous. They are quantitative symbols only. Amount or size merely indicates

[1] "Statistics * * * does not deal with a single homogeneous mass but with a complex body composed of multitudinous units differing in form and action one from the other; and it is with the complex not with the units that it is concerned." Bowley, A. L., *Elements of Statistics*, King, London, 1907, p. 262.

the presence or absence of a condition which is abstractly represented. Thus, units of length, width, and volume, conceived of in this manner, may be added, subtracted, or otherwise treated numerically as fancy dictates or necessity demands. This is done without any attention being paid to the units to which the symbols apply. They do not have to be adjusted to each purpose for which they are employed. For instance, a linear foot, as an abstract unit, is always 12 inches, a meter 39.37 inches, an American gallon 231 cubic inches. They may be combined with like units and converted into terms of each other without any serious inconvenience or risk of misunderstanding or confusion.

The same cannot be said of units of measurement dealt with in statistics. They are not abstract: they relate to some *thing* or *condition* which is concrete. Abstractly, all "ton-miles" are alike; concretely, they are different. While a ton is invariably a ton, and a mile a mile, all tons, except as to the one quality, weight, are not necessarily the same, nor are all miles, except as to the one quality, distance, always equivalent. One ton may be bulky, low-grade freight; another ton may be compact, high-grade freight. One may be the measure of a quantity of stovepipe elbows; the other, of a quantity of silks. Likewise, one mile may be of easy grade in a prairie; the other of heavy grade in mountainous tunnels. The conditions necessary to the movement of one ton one mile—the ton-mile—may be wholly dissimilar in spite of the common name which is assigned to the service. Statistical units have reference to things or attributes of things under different circumstances; combinations of them at will cannot be made. The fact is that in statistics, units of number, size, and frequency are dealt with not abstractly but concretely.

Units of measurement having to do with business, economic, and social affairs are often indefinite and general. By different people and under different circumstances, the same things are called by different names; or different things are called by the same name. Thinking and reasoning about them are

confused. People do not understand each other's use of terms. They do not use words and phrases having the same meaning or connotation, and, accordingly, interpret the same phenomenon in different ways or different phenomena in the same way.

Because of this and other facts, statistical measurements are often meaningless. Quantitative symbols are used to measure abundance or to indicate scarcity—more or less—but the symbols are attached to things which have different meanings. They are combined and averaged as though they were abstract. Confusion results when this is done. What is too often done is not to measure the frequencies of occurrence of the *same* thing, but of *different* things which are given the same name. An illustration involving the meaning of a unit will indicate the nature of the problem of statistical measurement.

If it is necessary to enumerate the number of "manufacturing establishments" in a given district, the definition of this unit will obviously be determined by the following, among other, conditions: (1) the meaning of "manufacturing" as distinct from trading, mercantile, transporting, agricultural, etc., pursuits; (2) the meaning of an "establishment." The definitions employed will depend upon the purpose in mind in using them. If it is to learn the number of such enterprises, and the test of identity is separate ownership, there may be many or few "establishments." If other tests, such as independent operation, unit housing, unit processes, unit management, contiguous location, etc., are imposed, then different numbers of "establishments" will be found. In such cases it is not enough to maintain that an establishment is an establishment. The identity, and therefore the number to be enumerated, depends upon the criteria which are used to distinguish them. The statistical process of grouping and combining individual instances into aggregates and of averaging them is impossible unless the units enumerated are identical in the particulars chosen as a basis for enumeration.

Another example of a somewhat different type may be given

in this connection. It is desired to determine the "industrial accident rate" in a given industry as a basis for fixing a scale of compensation for accidents. What is an "accident"? Obviously, the reason for compensation is personal injury with its attendant consequences, and it is the character of the injury which serves as a basis for enumeration. All injuries involving a loss of any time, howsoever slight, might be thought worthy of inclusion. But since compensation is the occasion for determining the number, only those injuries to which an *appreciable* loss of time is due should be included. What is an "appreciable" loss of time? To an individual who experiences the loss, such an amount might be any time, howsoever slight. To the employer, however, who advances the compensation, and to the public who finally bear it, a period of one or two weeks might be thought to be the minimum compensable period. But many trifling accidents may, in the aggregate, occasion a far greater loss of time than a single or a few serious ones. There would be no hesitancy about counting the serious ones, yet there might be doubt about including the minor ones. But it is precisely the latter which can most frequently be prevented, and about which information may be desired, because precautionary measures which involve little added cost to the employer, increased efficiency to the employee, and the gradual elimination of the occasion for compensation, may be taken to reduce them.

Moreover, by hypothesis, only *industrial accidents* are to be compensated. When accidents are enumerated for this purpose, self-inflicted injuries, as well as those occurring to workmen while not engaged in industrial operations, and when work done is not a proximate cause of injury, should be eliminated. Is "disease," contracted directly as a result of the conditions of industry, an "accident"? Surely it is an "injury," and if injury is the basis of compensation, ought not diseases contracted ir this way to be counted? If they are counted as an industrial injury (not "accidental," but characteristic or regular), how should instances involving impairment

of health, mental or physical ability, be considered? How long a period must elapse before a condition, the result of employment, ceases to be checked against such employment? What is an industrial accident for compensation purposes?

The unit of measurement, however, is the *rate of industrial accidents.* Not all occupations are equally hazardous, and to refer to industries the accidents occurring, irrespective of the occupations involved, is equivalent to assigning them to conditions which they cannot produce. Moreover, the number of accidents which occur is a function of the number of persons exposed to risks and the periods of exposure—the man-hours or man-days. In using reported accidents as a basis for compensation, care, therefore, must be taken to assign the results to conditions which produce them.

If the purpose in enumerating industrial accidents were, on the other hand, to measure the total amount of time lost through mental or physical injury, obviously *all* accidents and all diseases directly attributable to industry should be included. If the purpose were alone to secure information to be used as a basis for removing the conditions causing accidents, or for assigning responsibility for them as between employer and employee, machine and injured person, those which were trivial, from the point of view of the individual, would take equal rank with those which are called severe. What is an "industrial accident"?

Inquiries similar to the ones suggested respecting accidents must always be made and answered before the collection of primary, or the use and analysis of secondary data respecting any problem, is begun. It is not sufficient to study mere frequency, but frequency relating to the units chosen, and the units in their particular applications to the problems under consideration.

To formulate the purposes for which statistics are to be collected and used is the first step in statistical studies; rigidly and unmistakably to define the units of measurement in which the aggregates are expressed, and to adhere to them

throughout the process, is the second. The latter is governed by the former, as the former is determined by the latter. The two are reciprocal. Statistical units cannot be defined without regard to their purpose, and their purpose cannot be outlined with sufficient accuracy to be carried out without a clear notion of the units.

Probably enough has been said to bring to the reader's attention the meaning of units of measurement and the distinctions which must be made between the use of abstract concepts of mass or frequency in mathematical calculations and the use of the concepts in statistical studies. Statistics involves more than numbers and quantities. It is quantitative but has to do with more than numerical computations. *It is concerned, as has been said, with the processes and methods of formulating and testing conclusions from premises resting solely upon numerical bases.*

## II. Statistical Units of Measurement Classified and Described

It will be of assistance in understanding units of measurement to classify the different types and to describe their significance. Distinction should be made between (1) units of enumeration or estimation, and (2) units of analysis and interpretation.

The first are those in which measurements are made; the second are those in which they are compared. The first have primarily to do with collecting data; the second with comparing them.

### 1. UNITS OF ENUMERATION OR ESTIMATION

The units in which data are enumerated or estimated are either *simple* or *composite*. A simple[1] unit is one which is general in meaning, class differences only being distinguished. Examples of such units are the following: a farm, a ton, an

[1] See the discussion, *supra*, pp. 35-36.

accident, a strike, a lockout, an immigrant, a room, a street, a draft, a bill of exchange, a deposit, a novel, a citizen, etc. Such units are easily distinguished; they are mutually exclusive. No distinction is provided for degrees of similarity, but only for absolute differences. Such units have no limiting qualifications.

In contrast with simple units are those which are called *composite*.[1] Composite units are formed by adding to simple units a limiting or qualifying word or phrase, the effect of which is (1) to define more accurately the general concept, (2) to restrict the class which it names, and (3) to add to the difficulty of defining it. For instance, a "sale," as a simple unit, becomes composite by adding to it the limiting word "credit." The unit is now a "credit sale." To identify it, it is necessary not only to distinguish the condition of "sale" from that of purchase, for instance, but also to define what is meant by the term "credit." The simple unit "ton" becomes a composite unit by the addition of the word "freight." Similarly, an "accident" becomes an "industrial accident," etc.

To convert simple into composite units sometimes has the effect of changing the meaning and use, as well as the scope, of the term. For instance, the unit "room," in a survey conducted solely to determine the *size* of rooms in tenement buildings, might be defined as any portion of a house, habitually used as a place of abode, set off by walls with exits either closed or capable of being closed. To add to this unit the limiting word "sleeping" suggests so many considerations respecting light, ventilation, size in respect to number of occupants, time of occupancy, etc., as to alter materially the meaning attached to it when the counting is undertaken to determine size, but not size in connection with *use*.

To repeat, statistical processes are not confined to counting or combining abstract units, but have to do with those relating to particular circumstances and particular problems. For instance, it is desired to compare the illiteracy among Southern

[1] See the discussion, *supra*, p. 36.

European immigrants and the American negroes. It would be clearly an error to make this comparison until the meanings of "immigrant" and "negro" were definitely settled, until comparable sex and age classes were specified, and until the same or comparable tests for determining illiteracy were employed. Illiteracy tests established for immigrants may not have been the same as those used for negroes. The tests for the immigrants may not have been adjusted for the different age classes, nor determined according to standards characteristic of the New World. Moreover, they may have been influenced by the standards used to distinguish immigrants from non-immigrants.

The point emphasized is the necessity of reducing the conditions in every unit to a homogeneous basis. Those which are conflicting and overlapping cannot obtain. This applies particularly to cost accounting where it is necessary that cost data be reduced to their most elemental units. If composite or compound units are used, comparisons, except under the most favorable circumstances—circumstances which seldom if ever exist—are meaningless. This contention is brought out in the following citation relating to the use of cost units in New York City.

"An example of the weakness of the usual cost data is shown by the cost per square yard for certain paving work done by five different gangs under different foremen. I have in mind a single day's work for these gangs. The work to be done was identical yet the cost ranged from $1.11 per square yard to $1.89. This cost data was worthless on its face because it did not analyze the cost into the constituent elements. It accepted the *compound*[1] unit cost as final. By going back of the unit cost per square yard we find the reason for the difference in cost for doing the same thing under similar conditions. We base everything on *elemental*[1] cost data. By this is meant the unit cost of each element that enters into the performance of a thing as, for instance, the laying of a square yard of asphalt pavement. The fact that it cost only $1.70 for laying a square yard of asphalt pavement is absolutely useless and misleading unless we know all of the facts entering into the cost of laying the pavement." (Here follows a statement of thirty elements to

[1] Italics mine.

be considered in making such comparisons.) * * * "The fact is that one square yard of asphalt may be cheap at $2.00, while another square yard may be high priced at $1.00.

"Another trouble with *compound*[1] units cost data is that it compares entirely dissimilar things with each other. * * * The number of square yards to be done has a marked effect upon the unit cost per square yard and the conditions under which the work is done will have an even more marked effect."[2]

### 2. UNITS OF ANALYSIS AND OF INTERPRETATION

In contrast with units in which things or attributes of things are *named,* as for instance by the simple units "stores," "houses," "sales," or by the composite units "chain stores," "bond houses," "forced sales," are those in which things or attributes of things are *compared as well as named.* To compare things they must be placed in relation to each other. To do this requires the use of ratios, or coefficients[3] as they are sometimes called.

Comparisons may relate to time, to space, or to conditions in time or space. Illustrations of ratios or coefficients involving these points of view will serve to make the distinctions clear.

#### *(1) Ratios or Coefficients Relating to Time*

Sales of retail stores or the wages of working men may be *expressed* in dollars, but *related* to days, months, or years. If in comparing sales, the time unit *year* is taken, such a period may be unsuitable, because, in the different establishments, (1) there may be a seasonal element in one line of trade and not in another; (2) the goods sold may have different seasonal characteristics; (3) the sales in one may be spread over the

[1] Italics mine.

[2] Adamson, Tilden, "The Preparation of the Estimates and the Formulation of the Budget—The New York City Method," in *The Annals of the American Academy,* November, 1915, Whole No. 151, Vol. LXII, at pp. 253-255.

[3] See the discussion, *supra,* pp. 36-38.

entire period; in another, be crowded into a few months; (4) the beginning and close of the year may vary.

If wages of workingmen in different industries although *expressed* in dollars are related to days: that is, if the coefficient "dollars per day" is used, comparisons may be faulty because (1) the days are of unequal length, or (2) the number of days customarily worked in a year, for instance, is different.

Again, industrial accidents may be *expressed* by number or by severity, but *related* to years. Those occurring in different plants, however, within a given year, may vary because of (1) the number of days the plant operates; (2) the number of employes used, and the length of time they work; (3) the relative bazard of each occupation; (4) the different proportions of the total force engaged in the hazardous occupations.

### (2) *Ratios or Coefficients Relating to Space*

For different states the amounts of wheat raised during a given season or year may be *expressed* in bushels. They may be *related* to 100 square miles of territory, counties, farms, etc. The space units—the denominators of the different coefficients —may be unsuitable for comparing different yields because (1) not all square miles, counties, or farms produce wheat; (2) the counties and farms may be of different size; (3) different proportions of the square miles, counties, and farms may be used for wheat production.

Again, sales may be *expressed* in dollars, and *related* to hundreds of square feet of floor space. But (1) not all floor space is used for sales purposes; (2) the proportions of the total used for this purpose, in different establishments, vary; (3) the floor space is probably not uniformly placed with respect to floors, frontage, etc.; (4) the types of goods sold on different parts of tbe space used for the purpose vary in price, at a given time, and during different seasons of the year; (5) different grades and proportions of the same variety of goods are displayed, etc.

### *(3) Ratios or Coefficients Relating to Condition*

Deaths during a given period, or for a given area may be *expressed* in numbers. They may be *related* to the entire population or to the population of the same age and sex characteristics. If the first basis is used, the coefficient—deaths per 100,000 of population, for instance—is faulty, because (1) all elements of a population are not equally likely to die; (2) the age and sex characteristics of populations in the same place at different times, and in different places at the same time are not necessarily the same; (3) the proportions of the total deaths from different causes may vary from period to period, and from place to place; (4) epidemics of the same duration causing deaths may not be regular in their occurrence, universal in their appearance, nor equally deadly in their effect.

If deaths are related to populations of the same age and sex characteristics, some, but not all, of the limitations of the cruder bases are removed.

Again, total operating expenses of retail establishments may be *expressed* in dollars. The amounts may be *related* to $100 of sales. The coefficient would then become "total expenses per $100 of sales"—expenses constituting the numerator, and sales in hundreds the denominator of the ratio. But (1) all expenses do not have to do with sales; (2) both expenses and sales in different stores result from different types of services rendered and goods sold; (3) the proportions of the expenses and the sales, attributable to different sources, vary.

The turnover of retail merchandise during different periods for stores of different size, or with different location may be measured. The number of turns is secured by dividing the cost of merchandise sold by the amount of average inventory or stock on hand taken at cost price. That is, a coefficient is employed. Both the merchandise sold and the stock on hand are taken at cost price. To express the numerator in terms of cost and the denominator in terms of sales price is incor-

rect because (1) cost and sales bases are not identical; and (2) gross margins—the difference between the cost of goods and their sales price—may not be uniform for different types of goods, nor for different merchants.

These illustrations of coefficients or ratios relating to time, space, and condition will suffice to make the distinctions between them clear. They probably do not, however, make it plain why some coefficients are satisfactory and others unsatisfactory. This may be done by stating the general principle which should be followed in setting up all types of coefficients. The following are different ways of expressing the essential idea.

1. *Compare only those things or attributes of things which are alike or have common qualities.*

2. *"Always relate effects to the causes producing them."*

3. *The denominator in every coefficient should relate specifically to the condition named in the numerator.*

4. *"The numerator should be homogeneous and the denominator should be homogeneous, and each unit in the denominator should bear the same potential relation to the attributes of the units in the numerator."*

If these rules are not followed, comparisons break down. The result is that "crude" rather than "corrected" units are employed. The "crudity" may relate to a time, a space, or a condition factor, depending upon the type of unit which is used. To correct a coefficient is to follow the principle stated.

Comparisons relating to remote periods, widely separated places, or different conditions are always questionable.[1] Too

[1] The following cautions are of interest respecting the difficulties of comparing railway statistics in the United States and foreign countries: "Attention is called especially to the fact that the strict comparability of all the items throughout this bulletin is not assured, even by the greatest care in compilation. It would be an impossible task so to tabulate and adjust the railway statistics of a number of countries—differing from each other in so many respects—as to place them on a strictly comparable basis. Every attempt to present a comparison between statistics of different countries encounters practically insuperable obstacles to complete comparability. These spring from numerous differences in the classification of data, in the composition of accounts,

great care cannot be taken to make them legitimate. This is particularly true in the case of statistical comparisons, since they are numerical and seemingly exact. A numerical statement of a fact is often taken by the unwary and uninitiated, as sufficient proof of its absoluteness and finality, and is made to support predetermined conclusions or premises to which it has no relation. A rigid adherence in the collection of primary, and in the use of secondary data, to the principles here formulated respecting units, will help the reader to use statistical facts in a scientific manner.

and in the organization and character of the railway service. A few examples will illustrate the point.

"In most European countries the term 'freight,' as employed in the statistics of freight tonnage and freight revenue, includes a large part of such traffic as is carried by express companies in the United States. . . . A great part of such traffic is carried on fast freight trains along with what Americans designate 'package freight.' It is in most respects a part of the fast freight service, rather than an express service, as that is understood in the United States. Besides the question of expediency, is the impossibility—since both kinds of traffic are carried on the same freight trains—of determining for comparison on the train-mile basis the freight train-miles, in the American sense of the term, that would correspond to the revised tonnage and revenue statistics obtained by eliminating this sort of express traffic. By leaving this traffic in the tonnage and revenue statistics for freight, the data for each country are at least self-consistent.

"Differences in the character of the service affect the comparability of average receipts per passenger-mile and per ton-mile. In the case of the passenger service, practically all countries other than the United States and Canada offer a great variety of accommodations. And in those countries the cheaper accommodations, much inferior to that of the usual 'day coaches' here and in Canada, are far the more extensively used. As a result, the average revenue per passenger-mile is greatly reduced on account of the preponderance of traffic in the second, third, and even fourth classes. No allowance can be made for this difference by any adjustment . . .

"In the case of the freight service, the railways of the United States carry freight to a far greater extent in wholesale lots than in any other country except Canada. European countries, including England, cater to frequent, quick delivery of small shipments. The result is a more expensive service and a higher average charge. Furthermore, the average length of haul in the United States is . . . greater than in any other country. A comparison of the average receipts per ton-mile from the freight traffic as a whole in the United States and other countries is thus not a comparison of receipts for quite the same kind of service." "Comparative Railway Statistics, United States and Foreign Countries, 1912," *Bureau of Railway Economics, Consecutive No. 83, Miscellaneous Series No. 21, 1915*, Washington, D. C., pp. 7-8.

## III. Statistical Units of Presentation

Section II, immediately above, had to do with the different types of units in which statistical data are measured and compared. These were classified as (1) simple units, (2) composite units, and (3) ratios or coefficients. But data are not only *measured*, and *compared*; they are also *presented*. It is the various types of presentation units with which we are now concerned.

Age, for instance, may be measured to the nearest day, month, or year; size of city to the nearest thousand; and expense to the nearest dollar. Similarly, the composite units, selling expense, cost of merchandise sold, full-time salesmen, freight receipts, etc., may be recorded, counted, or estimated. Again, coefficients may be built up accurately or inaccurately. Simple and composite units have to do with enumeration or estimation; coefficients, with enumeration or estimation, *and* comparison. All of them involve measurements; they have nothing to do with the manner or way in which the measurements are presented.

Units of presentation are of three types: (1) time, (2) space, and (3) condition. For instance, the operating expenses of a group of retail meat stores may be measured to the nearest thousand and be presented by years, by location, by size and by nature of management; age may be measured to the nearest month, and be presented by years; heights may be measured to the nearest quarter of an inch, and presented in whole inches; live stock may be counted by farms, and be presented by states; railroad earnings may be secured by months and be presented by ten-year periods, etc.

*Units of presentation involving time are crude when the intervals used exceed those to which the measurements apply.* If earnings, for instance, are determined by months and show seasonal changes, accuracy is sacrificed by expressing them by years or groups of years.

*Units of presentation involving space are crude when the*

*areas used extend beyond those to which the measurements apply.* If population density in cities, for instance, is measured by blocks, and conditions vary in different parts of a city, the significance of this variation is lost by presenting the data by wards.

*Units of presentation involving condition are crude when the class limits used are so broad as not to reflect differences observed in measurement.* If costs of doing business, for instance, vary directly with volume of sales, then they should be presented in groups which will disclose this fact. Or, if costs of manufactured goods vary according to pattern of product, they should not be shown alone by entire output.

To convert crude into "corrected" units of presentation is to allow the peculiarities discovered in the measurements to be reflected in the way in which they are presented. To illustrate such a process: The costs of doing business are found to vary with location. This fact is discovered from the measurements themselves. How shall they be presented? Ideally, every variation should be indicated. Practically, this is impossible. Hence, areas are grouped, and cities classified according to size, the purpose being to select those units of presentation which will best reveal the peculiarities of the phenomena measured.

In general, the aim is to adopt that unit of time, place, or condition for presentation which will give the facts vitality and make them serve most fully the purposes for which they were collected or assembled. Statistics collected without a well-defined purpose are seldom of much value because of the lack of care in their preparation, and because of the absence of a controlling purpose in their presentation.

"Science has derived very little or no benefit from the miscellaneous collecting and grouping of facts without any previous notion of what they are likely to reveal. An investigation is usually made for the purpose of answering a definite question, or of verifying an anticipation. With some such end in view, with some principle by which the classification is guided, the result usually re-

veals not only what is looked for, but frequently still more fundamental characteristics * * *."[1]

Too frequently the groups into which facts are crowded are so broad, purposeless, and indefinite that whatever value the facts may have had as collected is lost by the failure to correlate the method of presentation with the purpose or function which they are to play. Thus death rates are tabulated by districts so large that correlation of deaths with their respective causes in detail is impossible. From an administrative point of view, such statistics are almost worthless. Similarly, causes of death are frequently tabulated in groups so broad and ill-defined as to make it impossible to single out from the groups the significant causes, and to use the statistics as a basis for a health crusade. Again, density of population—a common coefficient—is almost worthless when assigned to so large a population and so diverse conditions as those found in cities of appreciable size.[2] Density as a coefficient is significant where overcrowding is a problem. Not all sections of cities are capable of producing the unit of density assigned to the entire district, while in many sections the density is far greater than the single unit implies. In some districts density is of no significance; in others, it is precisely the unit which is most vital. The units of presentation should always be chosen with the thought in mind of making the statistics function.

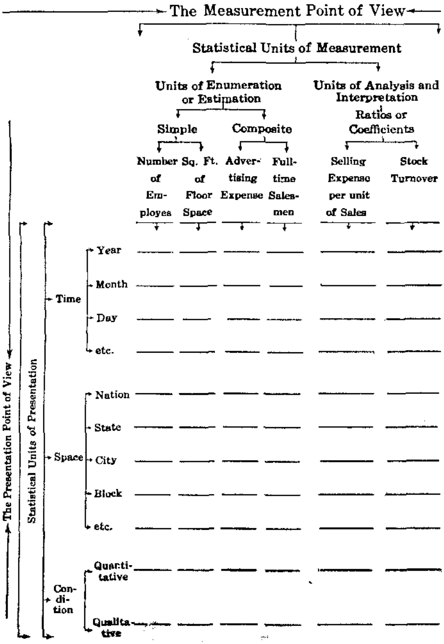
Taking an illustration from a more strictly economic field, a large part of our wage statistics, as presented for public consumption, suffers almost beyond redemption because they are reported as undifferentiated totals, as averages, or in groups so broad as to conceal the facts which they might otherwise reveal. The wages paid to a non-homogeneous class expressed as a total or as an average without classification is of little significance in throwing light on problems on which we need

[1] Cramer, Frank, *The Method of Darwin: A Study in Scientific Method*, McClurg, Chicago, 1896, p. 92.

[2] Cf. Bowley, A. L., *The Nature and Purpose of the Measurement of Social Phenomena*, King, London, 1915, pp. 40 ff.

light, such as the distribution of wealth, a sound basis for arbitration of wage disputes, standards for minimum wages, etc. The units of presentation are generally too broad; the facts are related to conditions which do not produce them.

### IV. Diagrammatic Scheme Illustrating Different Types of Statistical Units—Figure 1



## V. A Selected List of Units of Analysis and of Interpretation—Ratios or Coefficients

| The Unit | Formulæ Used to Compute the Units | "Crude" or "Corrected" |
|---|---|---|
| 1. Number of deaths per 100,000 of population | Deaths / Population (in 00,000's) | "Crude" |
| 2. Number of deaths from specific cause for specific age group per 1,000 of population of corresponding ages. | Deaths by cause in specified age group / Population (in 000's) of corresponding ages | "Corrected" |
| 3. Accident frequency rate | Number of accidents / Number of people employed (in 000's) | "Crude" |
| 4. Accident frequency rate | Number of accidents / Number of full-time workers (in 000's) | "Corrected" |
| 5. Sales per salesman | Sales / Number of salesmen | "Crude" |
| 6. Sales per full-time salesman | Sales / Number of full-time salesmen | "Corrected" |
| 7. Selling expense per hundred of sales | Selling expense / Sales (in 00's) | "Crude" |
| 8. Rate of stock turnover | Merchandise sold (at cost price) / Average stock (at sale price) | "Crude" |
| 9. Rate of stock turnover | Merchandise sold (at cost price) / Average stock (at cost price) | "Corrected" |
| 10. Rent per hundred of sales | Rent / Sales (in 00's) | "Crude" |
| 11. Rent per unit of floor space | Rent paid for first floor / Floor space rented (*in 00's*) of square feet on first floor | "Corrected" |
| 12. Working capital ratio | Total current assets / Total current liabilities | "Corrected" |
| 13. Turnover of accounts receivable | Average amount of accounts receivable / Average daily sales on account | "Corrected" |
| 14. Net ton-miles per loaded car-mile | Net ton-miles (in 000's) / Loaded car-miles (in 000's) | "Crude" |
| 15. Net ton-miles per loaded car-mile | Net ton-miles (in 000's) of specific freight / *Loaded car-miles (in 000's)* of specified freight | "Corrected" |

Why some of these units are called "crude" and others "corrected" the reader should be able to determine on the basis of the above discussion.

## VI. Rules for the Use of Statistical Units of Measurement and of Presentation

### 1. units of measurement

(1) Refer all units of measurement to the conditions which produce them. Make them homogeneous, suited to the purposes for which they are employed, and use them with consistency and integrity.

(2) Define clearly and fully all units which are used. Certain corollaries follow from this general rule:

a. Study problems in all their aspects before defining the units. Anticipate so far as is possible the difficulties to be encountered, and make provision, if possible, for others not foreseen.

b. *Define all units in the light of the intelligence of the* informants and the character of the data to which they apply.

c. Make all definitions in such a form that exceptions will be readily detected, misunderstanding of terms difficult, and employment ready, and in terms and form characteristically employed.

d. Establish a logical basis for all definitions.

e. Avoid substantive or descriptive units when direct ones are available.

(3) Appreciate the fact that statistics should be viewed functionally, and that a main source of error is in the units which are used in collecting and assembling data.

### 2. units of presentation

(1) Avoid "crude" whenever "corrected" units may be used.

(2) Seek to have units of presentation reflect the characteristics of data which are discovered in their measurement.

(3) Choose those units which are suited to the needs and purposes of the consumers to whom the statistics are presented.

## References

BOWLEY, A. L., *The Nature and Purpose of the Measurement of Social Phenomena*, King, London, 1915, pp. 29-97.

BOWLEY, A. L., "The Improvement of Official Statistics," *The Journal of the Royal Statistical Society*, *1908*, Vol. 71, pp. 461-469, on "The Nature and Condition of Statistical Measurement."

WATKINS, G. P., "Statistical Units," in *Quarterly Journal of Economics*, Vol. XXVI, pp. 673-702.

ŽIŽEK, FRANZ, *Statistical Averages* (translated by W. M. Persons), Holt, New York, 1913, pp. 25-33.

# CHAPTER V

# PURPOSES OF A STATISTICAL STUDY OF WAGES, UNITS OF MEASUREMENT, SOURCES OF DATA, SCHEDULE FORMS—ILLUSTRATIONS OF METHODS

## I. THE PROBLEM IN THE STUDY OF WAGES STATED

### 1. INTRODUCTION

IN the preceding chapters emphasis has been placed upon the logical order in statistical studies—(1) deciding upon the merits of the statistical approach, (2) outlining fully the purposes of study, (3) defining the units, and (4) assembling secondary and collecting primary data. The relations between these various steps are concretely illustrated in this chapter in a study of wages.

Much is now being written and spoken on the topic of wages. Socialists are condemning the "wage" system; social workers and those interested in ameliorating the condition of the poor are constantly urging the payment of a "living" or of a "minimum" wage. Wages is the bone of contention in industrial disputes, and by some is thought to be the ultimate source of all our industrial ills. Efficiency advocates are studying various methods of wage payment in an attempt to harmonize the principles of industrial efficiency with the interests of employes and thereby to enlist their support in having them adopted. Others are testing the level of wages in terms of their purchasing power either to measure their trend or to demonstrate their reasonableness. Still others are attempting to adjust to an increased nominal wage scale the prices charged for commodities and services in the hope of "making both ends

meet." To employes, wages are too low; to employers, they are too high. To one, they are income, to the other, costs. The importance of the subject in all its vagaries is sufficient reason for choosing it in order to illustrate certain principles of statistical methods.

It has been thought best to approach the problem from the standpoint of a public bureau collecting data from many employers, rather than from the standpoint of a single employer assembling wage data in his own establishment. The first approach, in a sense, includes the second, inasmuch as each employer must organize the material in his own plant before filling out the schedule for the collecting bureau. Moreover, employers are always interested in the wages their competitors are paying, and the only available sources for the necessary facts are the reports of public bureaus. They are likewise interested in the collection process, for only by a full knowledge of it are they in a position to know the meaning of collected data. The finished product is the basis for any comparisons which they may desire to make, and consequently its scope, merits, and demerits must be known.

When employers deal with their employes in matters affecting wage disputes, they need information on competitive wage scales; when they are concerned with their position in industry or trade, they need to know not only their own but also their competitors' labor costs.

There is another reason for approaching the problem from the point of view of an outsider. Units of measurement and types of reports are generally standardized within individual establishments. As between establishments, however, they differ considerably. For this reason, wage comparisons are often of little value, although they are given much weight, and it is the dangers involved in making them which are here given particular attention. These are traceable to (1) inaccurately and loosely defined units of measurement, (2) unrepresentative, biased, and crudely tabulated data, and to (3) the failure to understand what is involved in a statistical comparison. In

order to use statistics with discrimination and integrity, it is necessary to have a knowledge of their source, of the interpretation given to the original entries, of the groups and combinations into which they are thrown, etc. It is with these thoughts in mind that so much attention, in the preceding chapter, has been given to units, and that in this one the collection process for a concrete problem is discussed from beginning to end.

### 2. CHARACTERISTIC CONFUSIONS IN THE USE OF THE TERM "WAGES"

The meaning of the term "wages" in current discussions is generally clear from the context in which it is used. When the term is employed statistically, however, its various uses frequently cause misunderstanding and confusion. Wages and earnings are often used synonymously without any seeming appreciation of their differences. Wages and wage-rates, nominal or money rates and real wages are used interchangeably, or at least without clear distinction of the differences involved and the conditions upon which they rest. The term "salaries," as contrasted with wages, is used to distinguish large and regular from small and precarious incomes, notwithstanding the fact that the bases chosen are in part illogical when income as salary is less than income as wages. Moreover, the criteria by which the two are distinguished are not standardized; the rules set up are not always strictly adhered to and statistical studies, based upon current distinctions or in violation of them, sometimes lead to grotesque conclusions. The necessity of relating facts to the conditions producing them, and of making comparisons involving considerations of time, space, or condition legitimate, are constantly being violated.

The reasons for and types of confusion in the use of this expression will more clearly be seen by studying various purposes for which one would wish statistical information on wages, and by defining the limits of the term as used for these purposes. No attempt is made to cover all, but only those

purposes which bring out the peculiar statistical difficulties to which it is desired to call attention.

### 3. BASES FOR A DEFINITION OF WAGES

Wages are defined in current economic discussions as "the income received on account of labor performed,"[1] "the price of labor hired and employed by an *entrepreneur*";[2] or as including "all earnings assigned to men for their work, from lowest piece wages to highest annual salaries and 'wages of management.'"[3] In a still different sense the term is used to indicate "the share of the annual product or national dividend which goes as a reward to labor, as distinct from the remuneration received by capital in its various forms."[4] The term thus defined is too indefinite for statistical use, yet the distinctions suggest the differences to which it is desired to call attention. The first suggests property as contrasted with service income,[5] but does not distingush money income from real income nor salaries from wages. The distinction between the wage system and other possible methods of service remuneration is reflected in the second, while the last calls attention to a use restricted to economic theory—namely, that of distinguishing the reward of labor as contrasted with the reward of landlords and capitalists.

A number of distinctions must be made in order to use the term in statistical studies. Wage-rates must be distinguished from earnings; nominal rates from real rates; and earnings from labor—wages—from earnings from all sources including returns from investments, rents, etc. It is necessary also to distinguish wage-rates from salary-rates, and wages (wage-rates times the period for which paid), from salaries (salary-rates times the period for which paid). In converting

[1] Johnson, A. S., *Introduction to Economics*, p. 152.
[2] Gide, Chas., *Principles of Political Economy* (Second American Edition), p. 487.
[3] Seager, H. R., *Principles of Economics*, p. 244.
[4] Webster, *New International Dictionary*.
[5] See Nearing, Scott, *Income*, Macmillan, 1915.

wage-rates into wages the former must be increased by the money equivalent of concessions and perquisites and decreased by the money equivalent of time lost for which no compensation is received. Money wages must be clearly differentiated from real wages, or "the purchasing power of nominal wages measured by a constant standard." When computing real wages and making allowance for concessions, perquisites, payments in kind, and unemployment, the nominal money equivalent must be reduced to its purchasing power and added to or subtracted from, as the case demands, the money wages similarly reduced.

### 4. WAGES DEFINED

The term "wages," therefore, will be used to suggest various concepts but always with the following meanings:

By wages, when used alone, are meant earnings in money or its equivalent because of manual, mechanical, or clerical labor service, paid according to a stipulated scale, at frequent intervals, and under conditions which make it customary to make deductions for short periods of time lost. This definition does not admit of the term being used to cover labor's "share" in contrast with the shares of capital and land in distribution.

By wage-rates are meant the predetermined rates at which manual, mechanical, or clerical labor service is remunerated. Wage-rates multiplied by the period for which paid equal wages as defined above.

By salaries are meant earnings in money or its equivalent because of responsible, supervisory, or directive labor service, paid according to a stipulated scale at infrequent intervals and under conditions which make it customary not to make deductions for short periods of time lost.

By salary-rates are meant the predetermined rates at which responsible, supervisory, or directive labor service is remunerated. Salary-rates multiplied by the period for which paid equal salaries as defined above.

By *earnings*, when used alone, are meant money incomes or their equivalents received for labor services, without distinction between wages and salaries. The same term, in order to include other income than that regularly received from labor service, must be accompanied by a limiting expression.

By *real wages are meant the equivalents of money wages* in economic goods measured in terms of a constant standard of value.

Some of the purposes for which statistical studies of wages, as currently understood, may be undertaken, and the meaning *which the expression must have in each case will now* be discussed.

## 5. STUDIES OF WAGES AND THE USES OF TERMS

If the purpose of study were to approximate the effect which trade unions have upon wages, one would be inclined at first to restrict the study to wage-rates, since minimum scales are determined by unions in bargaining with employers. Union figures on wages are invariably quoted as rates and are usually nominal and minimal. The actual rate received is frequently higher than the specified minimum; in some cases it may be even lower. If by wages are meant earnings from manual, mechanical, or clerical labor service, then the effect of union activities on employment would have to be considered. Wage-rates may remain the same and still wages be materially affected. This fact introduces other difficulties. Are unemployment, strike, and other benefits to be considered offsets to wage losses, or are they considered to be counterbalanced by increased dues necessary to replenish depleted unemployment, strike, and sickness funds? Union activities may seriously affect wages but have no influence on earnings from other sources. Wages, therefore, must be distinguished from earnings, if the latter are meant to include earnings from other than labor services.

When "minimum" wages are discussed, wages, undoubtedly,

are understood to mean rates, since employers are not compelled to hire labor but only to pay at least the stipulated minimum to those employed.[1] On the other hand, when the term "living" wage is used, reference is not so much to the rate of wages nor even to wages alone from labor service, as to earnings from all sources under the conditions possible for the persons affected. Undoubtedly, earnings from other sources than labor service, in the cases of those to whom the receipt of a living wage is a problem, are almost negligible, yet the term "income" is more suitable than the term "wages" to describe this condition.

In comparing wages for manual, mechanical, and clerical labor service by industries, occupation, districts, etc., it is necessary to use wage-rates instead of wages, since only the former are generally available. It is next to impossible to trace individuals from industry to industry and to approximate, with any degree of accuracy for an extended period, the extent of unemployment, the amount of overtime worked, etc.[2] It is doubtful if anything better than classified rates are procured by statistical bureaus which ask for earnings. The rates as quoted by trade union sources are always minimal and nominal and, therefore, are of limited significance in determining the economic status of the groups concerned. Those secured from employers are for a limited period—generally a week, except in intensive studies—and are not a satisfactory measure of earnings from labor service. Wages instead of rates are necessary for this purpose. The same fact applies in studies relating wages to efficiency, to sex, to nationality, to

[1] The order on minimum wages in the brush-making industry in Massachusetts specifically takes account of the rates to be paid. "Assuming an average scale of 50 hours and regular employment" (a rather violent assumption) "this rate (15½¢) would yield earnings of $7.75." Quoted from "Estimates of a Living Wage for Female Workers," by Charles E. Persons, in *Publications of the American Statistical Association*, June, 1915, p. 577.

[2] For the difficulties involved even in an intensive study, see "Wages and Regularity of Employment in the Cloak, Suit, and Skirt Industry," etc., *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 147, June, 1914, pp. 14, 41, 42, 56.

length of service, etc. Wage-rates are the only data generally available and, of course, should be used as such.

If the determination of the trend of wages is the problem to be studied, wages may mean a number of things. Wage-rates, or earnings in the broad or in the narrow sense, may be considered. Study may extend to nominal or money wages or to real wages, and may include not only wage labor but salaried labor as well. If the trend of real wages—"the purchasing power of nominal wages measured by a constant standard"—is the object of study, rates and not earnings must be used, since it is only the former of which we are in possession, or which we may secure with reasonable accuracy on an adequate scale. Homogeneous wage groups must also be used. Moreover, a logical basis for the inclusion or exclusion of salaries must be established, care being exercised that the basis of distinction is followed throughout the entire period. Nothing is here said about the price index used in making the conversion of wage-rates into current prices or of the peculiar difficulties in adjusting the index to the classes of labor to which the comparison applies.

If the purpose of a study of wages were to determine from the producers' standpoint the relative costs involved in labor service, as contrasted with rents or interest, obviously, rates of wages in the narrow sense used above would be too exclusive a category. Distinctions between salaries and wages would be unnecessary, since the purpose is merely to determine production costs assignable to labor as distinct from land and capital. If the approach to the same problem were made from the social viewpoint, it would be necessary to distinguish between wages and salaries, and on grounds other than those generally followed, inasmuch as those are frequently illogical and indeterminate. Merely to call one group salary receivers and another group wage receivers results in confusion when the economic conditions of both are similar, and when criteria for determining the status of one apply with equal force to the status of the other. There would be the same

reasons for accurately defining salaries as for defining wages. The bases for the definitions should be factors of importance in the study in which the units are used. It is inappropriate to contend that the conditions according to which the units are defined change with each purpose and, therefore, that such units are unsuitable for statistical uses. The premise is valid, but the conclusion does not follow. Such a claim only serves to bring more forcefully to mind a fact already considered, namely, that while abstract measures of numerical frequence are employed in statistical studies, they are not used abstractly but are applied to units the limits and terms of which are conditioned by the uses to which they are put.

## II. The Relation of the Problem as Outlined to Statistics of Wages

The preceding discussion has served in a general way to show the necessity of accurately defining units of measurement in connection with the purposes of statistical studies, and to emphasize the necessary points of distinction in the use of such a word as "wages," but it has probably not related, with sufficient closeness, the subject to actual statistical data and suggested the problems by which one is confronted in using wage data possible of collection or currently collected. This closer relation we shall now establish by indicating the sources for primary wage data, by discussing the difficulties experienced in their collection, by describing the types of secondary data currently collected, and finally by constructing wage schedules to be used in connection with a concrete problem.

### 1. SOURCES FOR PRIMARY DATA IN WAGE STUDIES

#### *(1) Primary Data Directly Applicable to Studies of Wages*

Primary data in the study of wages may emanate from four sources. Those secured from employes, from employers, and from union officials are directly applicable; while those from

institutions such as banks, building and loan associations, insurance companies, lodges, etc., are only indirectly applicable.

### a. Data from Employes

Data on wage-rates; hours of work (nominal and actual); the amount of unemployment by cause; the methods and frequency of wage payment; earnings from labor and from other sources; perquisites in the forms of bonuses, benefits, profits; penalties, fines, forfeits, union dues; budgetary expenditures, and facts relating to age, sex, nationality, occupation, training, length of service, previous wages, etc., may be secured in whole or in part, satisfactorily or unsatisfactorily, from individual employes, in proportion as informants are wise or ignorant, truthful or deceitful, willing or unwilling to aid, and in proportion as the statistical organization used is well- or ill-adapted for the purpose in mind. It is impossible to summarize in a single sentence the success attainable in securing data on wages or on any other topic directly from individuals involved. Frequently, the costs are prohibitive; in other cases, where cost is not an insuperable barrier, the types of individuals dealt with and the character of the information desired make this approach impossible. The generalization, however, is hazarded that data collected from a source where personal supervision or intimate checking is impossible are likely to possess serious limitations respecting all topics which in any way call for discrimination, for the exercise of judgment or the use of records, etc., on the part of the informant, or in which the personal equation enters to an appreciable extent. The discussion, in Chapter III, of *The Collection Process* is particularly applicable in this connection.

### b. Data from Employers

Much the same types of wage data as those listed above are theoretically obtainable from employers, and the chances are much greater that they will be free from error since less igno-

rant groups, recorded facts, impersonal relations, etc., are dealt with. The facts, however, are of a somewhat different sort and rarely apply to an extended period. The best that can be done in most cases is to secure cross-section views at widely separated intervals. Moreover, for the most part, classes and not individuals are considered. These may or may not be homogeneous, and in this respect are much less desirable statistical units than are individuals. From this source, with an adequate statistical organization, and with sufficient sanction, the total wage bill, time- and piece-rates, by occupations and processes, classified wage-rates, perquisites allowed and penalties assessed, and the number of employes classified by sex, age, and time of employment, etc., are theoretically available. The facts regularly secured on an extended scale and available for use are discussed below.

### c. Data from Trade and Labor Unions

In many respects the records of trade and labor unions are satisfactory sources for wage data. Theoretically, nominal time- and piece-rates for regular, for overtime, and for Sunday and holiday labor; nominal hours per day and per week; benefits allowed, classified by the amounts paid, by purposes, by duration, etc.; union dues; numbers unemployed, classified as to causes, and wage losses, etc., are available from this source. The data, however, may have serious limitations. Frequently, the desire to make out a case is held to be sufficient cause for furnishing defective returns or for withholding information. In many instances the inquiries addressed to union officials concern matters about which they can have but the most inadequate and superficial knowledge, and yet they are urged to give positive, negative, or numerical answers with few or no opportunities being offered for explanations. In some instances, undoubtedly, sincere efforts are made to state the truth as nearly as it can be determined; in other instances, no such care is exercised. The value which data from this source

possess is to a large degree determined by the scrutiny to which they are subjected hy collecting agents.

The limitations, however, are not always to be attributed to errors in reporting nor to incomplete returns. Frequently, they result from misusing and assigning finality to figures at best but estimates, from ignoring the specific advice of collecting agents, and from violating the fundamental principles of statistical methods. The same result, however, may occur respecting data drawn from the most acceptable sources. Statistical facts will be cited to prove contentions with which they have no connection and will be distorted and misapplied so long as people have hobbies, lack integrity, or are ignorant of the functions, limitations, and purposes of statistical data and legitimate ways of using them.

It will be noted that data on wages from unions are restricted to nominal rates and to union members. These are serious limitations where wages or earnings are sought and where non-union labor is involved. Such data are of little value in discussions of minimum wages, living wages, or other topics in which light is desired primarily concerning unskilled labor.

### *(2) Data Indirectly Applicable to Studies of Wages*

Facts which contribute indirectly to a knowledge of wages and wage conditions may be gleaned from a study of the increase or decrease of savings, the number of depositors in savings institutions and the average deposit, the size of employers' payrolls, the activities of building and loan associations, the growth or decline of fraternal insurance, the increase or decrease of union membership, etc. In most respects their connection with the topic is remote and contingent. They are at best suggestive and corroboratory and should be used with extreme caution, cognizance being taken of the roundaboutness of their application, the potency of other contributing causes to produce the effects shown, the interrelation of economic phenomena, etc.

Having sketched the types of wage data theoretically available, their sources, and the difficulties in securing and the dangers in using them, we may now briefly enumerate the types currently collected with their sources and some of their peculiarities. No attempt is made to describe or criticize fully or even to enumerate all forms regularly and irregularly collected in the United States. This has been done in a general way by others.[1] Moreover, such a treatment is not germane to our immediate purpose.

## 2. TYPES OF SECONDARY WAGE DATA

Secondary data on wages collected from the chief primary sources are available in many forms. They appear in public and private reports, issued on the basis of data furnished by wage earners, employers, and unions. Some reports appear regularly, some irregularly; some are restricted to the single topic, while others bear upon it only indirectly. Some are monographs on special topics, while others are exhaustive independent surveys.

### *(1) Secondary Data Directly Applicable to Studies of Wages*[2]

#### a. Data from Employes

Wage studies, in which the material is drawn from individuals alone, are made primarily in connection with cost of living

[1] Nearing, Scott, *Income*, Chapter II, pp. 18-52, New York, 1915; Streightoff, F. H., *The Distribution of Incomes in the U. S.*, Columbia University Studies, Vol. LII, No. 2, 1912.

[2] In this revision, it is thought not to be necessary to bring up to date the descriptive details in this chapter. The types of wage data which are collected, the manner in which collection is made, and the way in which the data are published are constantly changing. The details furnished, while not necessarily complete nor accurate for 1925, are sufficiently suggestive of the conditions which obtain. This is all they are intended to be. An introductory text on statistical methods is not intended to be an encyclopedia of statistical practices.

studies, such as those of Chapin [1] and Mrs. More [2] in America; Rowntree [3] and Booth [4] in England; or as a condition of the administration of labor laws, such as those on compensation for industrial accidents. Those of the first type generally apply to limited territories and restricted groups, cover only a relatively short period, and are made in connection with or are designed to throw light upon budgetary matters. In those of the second group, wage data are subsidiary to the main purpose of study, are restricted to definite classes, are not collected simultaneously for all groups, in some instances are semi-confidential, and are generally too meager to be conclusive respecting either ruling wage-rates or wages. Hence, they are not generally published except in summary form along with accident and other data.[5] They are, however, of excellent quality, because of the purposes for which collected, and in the course of time when they have been sufficiently accumulated will undoubtedly furnish material for thorough and comprehensive wage studies.

Studies on wages from material drawn directly from employes are published *only at irregular* intervals and cannot wholly be relied upon for current information. Those associated with budgetary matters refer invariably to wages or to earnings; those arising out of the administration of labor laws always relate to rates of wages. Those of the first class are important in calling attention to low wages in certain industries, in certain districts, for limited groups, and are indispensable in the determination of minimum and living wage standards, but are inadequate for comparing wages by indus-

[1] Chapin, Robert C., *The Standard of Living Among Workingmen's Families in New York City,* Charities Publication Committee, New York, 1909.

[2] More, L. B., *Wage Earners' Budgets,* New York, 1907.

[3] Rowntree, B. Seebohm, *Poverty; A Study of Town Life,* London, 1906.

[4] Booth, Charles, *Life and Labor of the People,* London, 1891.

[5] The brief tables on wages in "First Annual Report of the Industrial Accident Board," *Massachusetts Industrial Accident Board,* Boston, 1914, and in "Report No. 4" on "Industrial Accidents in Ohio, January 1 to June 30, 1914," by *The Industrial Commission of Ohio,* Columbus, Ohio, 1915, are illustrative.

tries, by localities, and over long periods. Neither do they furnish material for measuring the trend of wages. Those of the second class may be used to correlate wage losses and amounts of compensation for accidents, but at present are in the main superficial and restricted studies, serving no other purpose than that of a record of wage data collected on accident schedules.

### b. Data from Employers

The statistical matter relating to wages and wage conditions reported and published by regularly constituted statistical bureaus, by special commissions, and by individual investigators, may be divided into two groups; those directly related and those remotely connected with the topic.

#### (a) Material Directly Related to Wages

Direct material relates, first, to the total wage bill paid, and second, to classified wage-rates. The United States Bureau of the Census publishes at decennial and at certain intercensal periods the total salary and wage payments, made during the year to which the census applies, to salaried officers, to superintendents and managers, to clerks, stenographers, and other salaried employes, and to wage earners including piece workers in manufacturing and mining industries. The Interstate Commerce Commission publishes monthly the amounts paid to railroad employes classified into one hundred and forty-eight classes. The same commission publishes for express companies the wages and salaries of employes in the "traffic," "transportation," and "general expense" divisions. A few state bureaus of statistics and labor, particularly those in Massachusetts, New Jersey, and Ohio, collect and publish, as part of their manufacturing censuses, the total compensation for labor services classified as salaries and wages. The schedule[1]

[1] *Bureau of Statistics of Labor and Industries.*

used by New Jersey calls for the "total amount in wages paid during the year," and instructs informants that "only wages paid to wage earners actually employed" in an establishment or in "erecting or placing its products elsewhere" should be included. Salaries of managers, bookkeepers, salesmen, etc., are to be omitted. The schedule [1] to manufacturers used by Massachusetts asks for the "total wages (paid during the year to wage earners only)," and instructs the informants to omit "salaries of agents, managers, bookkeepers, clerks, salesmen, and others of this class." The schedule [2] used by Ohio contains essentially the same questions and provides for the same omissions, except that salespeople are divided into two groups, traveling and non-traveling.

Classified weekly wage-rates are collected and published for manufacturing enterprises in a number of states, but most satisfactorily in Massachusetts, New Jersey, and Ohio. In *those instances the data are taken from payrolls.* Massachusetts and Ohio in their schedules ask specifically for weekly rates, while New Jersey apparently desires weekly earnings.[3] Massachusetts and New Jersey supplement their schedules by field agents. Ohio is able to dispense with these in connection with her wage studies, inasmuch as, in the administration of her compensation law, she secures the audited payrolls of all employers subject to the law. It is not likely, under these conditions, that employers affected by the law in both respects will furnish incorrect returns.

The most exhaustive study of classified wage-rates for the United States is that on *Employees and Wages* made by the Census Bureau in 1903 under the direction of Professor Davis R. Dewey, and known as the "Dewey Report." The data refer to the years 1890 and 1900, apply to thirty-three industries, but include only a limited number of establish-

[1] *The Bureau of Statistics, Division of Manufacturers.*

[2] *The Industrial Commission.* It is not quite correct to speak of a "Manufacturing" census in the case of Ohio.

[3] The data are published as "earnings" but undoubtedly are rates.

ments in each industry. Wages of 103,453 employes in 1890, and of 160,859 in 1900 were tabulated in detailed groups. While the study is exhaustive in scope and unique in method it is not of current interest and must be passed over with brief mention.

The United States Bureau of Labor Statistics publishes from time to time special studies on wages and hours in different industries. These are always of interest. Indeed, this Bureau is the source from which most satisfactory data may be expected.

### (b) Material Indirectly Related to Wages

The material indirectly bearing upon wages may be classified under two heads, first, actual or average number of employes by months, and second, the time which plants operate during the year.

The United States Bureau of the Census publishes for manufacturing and mining industries the number of wage earners, including piece-workers, as per payrolls or time records, on the fifteenth day of each month for the periods covered by its reports. No distinctions are made for age and sex classes. New Jersey, as a part of her manufacturing census, publishes the "number of persons employed"[1] during each month of the year for which study is made, classified by sex for those sixteen years of age and over, but without sex classification for children under sixteen. Massachusetts publishes the average[2] number of wage earners during each month for

[1] Neither the instructions to informants nor the schedules define this number. Whether it is to be the average force computed on the basis of twenty-six, thirty, or thirty-one days, to be the normal force during the period, or the number of separate individuals to whom employment was given during each month, we are not told. It conceivably might be any one of them, carefully computed, but more likely it is a rough average representing nothing better than an estimate.

[2] The use of an average in this case seems unnecessary and somewhat to lessen the value of the figures in computing the deviations from month to month, with the purpose of throwing light on the seasonal character of employment. There seems no sufficient reason why the exact number, as required by Ohio, and others, should not be called for.

males and females separately but without age classification. She likewise publishes the number of wage earners eighteen years of age and over and under eighteen years of age classified by sex on the thirteenth [1] day of December as per payroll. Ohio requires employers to report the number of wage earners employed on the fifteenth day of each month as per payroll, classified by sex but not by age.

Ohio, likewise, requires employers to report the number of full days that plants are in operation and idle during the year, the former including part-time days reduced to a full-time basis and the latter not including Sundays and holidays unless plants normally operate on these days. The number of hours normally worked per full day or shift and per full week is also required to be reported. In Massachusetts the number of days in operation and idle is included in the manufacturing schedule and published in this form. Informants are specifically reminded that the working year is composed of a stated number of days and that the sum of the days reported, not counting Sundays and holidays, should total to this number. In New Jersey, data are published for manufacturing establishments on the number of days in operation, the normal number of hours per day, the normal number of hours per week, and the total number of hours extra time during the year in which establishments operate. The Bureau of the Census publishes like figures on the number of days manufacturing and mining establishments are in operation during the year and the number of hours normally worked by wage earners per shift and per week. Respecting the latter topic, informants are instructed that "all that is desired to know is the practice generally prevailing in respect to the hours of labor of employes."

### c. Data from Trade and Labor Unions

The wage data regularly collected from union sources by statistical bureaus refer to nominal (minimum) time- and

[1] This is the date indicated in the schedule for 1913.

piece-rates, nominal (maximum) hours per day or per week, causes and extent of unemployment, number and duration of strikes, etc. In this descriptive part of the chapter it will suffice, in view of what has been said above, briefly to describe the statistical activities of the United States Bureau of Labor Statistics, of the Department of Labor of the State of New York, and of the Bureau of Statistics of Massachusetts, respecting union wage conditions.

The United States Bureau of Labor Statistics has published the union scales of wages and hours of labor for the principal mechanical trades, for the largest cities of the United States for the period 1907 to date. The report for 1913 covers the forty industrial cities located in thirty-two states for which the Bureau publishes retail price statistics. Union scales for both wage-rates and weekly hours are followed, but such scales fix the limits in only one direction. Minimum wage-rates are established below which members of unions will not as a rule work, and maximum hours beyond which they will not work at regular rates of pay. In certain cities and trades, workmen are paid more than the union scale and work regularly less than the scale of hours. However, the Bureau takes no cognizance of these conditions. All wage-rates are reduced to an hourly basis, and for all the trades for which the Bureau has figures, relative or index numbers are computed for both wage-rates and hours for the years 1907 to 1913. The data are collected by special agents in personal visits to union business agents and secretaries, and the wage scales, written agreements, and the trade union records consulted wherever available.[1]

Statistics of unions and their membership were first collected by New York State in 1894 and 1895. Since 1897 such statistics have been regularly published. Information is now collected semi-annually from all unions, in part by schedule and

[1] A similar study, in co-operation with the United States Bureau of Labor Statistics, is made by the Industrial Commission of Ohio and applies to all the larger cities in the state.

in part by field agents. Schedules relate to membership and idleness, to hours of work, to new trade agreements, to changes in the rates of wages, and to rates of wages of time workers. The amount of unemployment is reported under six specific and one miscellaneous head; lack of work, lack of material, the weather, strikes or lockouts, sickness or accident, old age, and miscellaneous. The data apply to the sexes separately and to the end of March or September as the case might be. The regular hours of work for Saturday, Sunday, and other days, and the total per week by branches of trades and for the sexes separately are included. Changes in hours, with those before and after each change, and number of persons affected are also requested. Respecting rates of wages, information is secured on the rates before and after changes, the number of members affected, and the estimated weekly earnings before and after changes in the case of piece workers. Schedules respecting wage-rates of time workers relate to each branch or grade of work, to the working hours per day for the specified rates, and to the number of members by sex receiving them. Other inquiries of less significance and certain modifications of these are also included. It is unnecessary for our present purposes to supply more details.

The schedule is a model in technique; the questions are vital, clearly stated, and well arranged. It is mailed to union secretaries, ten days are given for answering, and delinquents are visited by field agents of the Bureau. Approximately 50 per cent of the schedules are sent in by mail and 50 per cent "fielded."

The published material is issued in two series: one called "Series on Employment" and the other "Series on Labor Organization." The first shows the amount of unemployment by cause, by months, and includes summaries for years by industries and by detailed trade groups. The issuance of a letter on the state of the labor market based upon monthly returns from the larger unions is also a regular feature of the Bureau's activities. The second series relates to the number

and membership of unions classified so as to show data by industries, by trades, by localities, etc.

This account of the New York Bureau's activities respecting union wages and conditions, although brief and sketchy, is probably adequate to reveal in a general way the types of data collected and the manner of securing them. Neither the schedules nor the methods of tabulation are open to severe criticism. The only criticism which might be offered is that the facts are supplied by unions. Essentially the same facts, but in a different form, respecting wages, hours, and unemployment, are available from employers and the probabilities are that they are more accurate when so returned than are those furnished by unions in spite of the care exercised to correct the errors. Employers are subject to state supervision in many respects, the statistical machinery is adjusted to this source of information, and the reporting of facts may be required legally. Unions are not compelled to report nor are they punished for withholding or distorting the matter supplied. In one respect, however, it seems necessary to deal with unions as units. Public and private boards of arbitration require *union* scales of wage-rates and hours as bases for making awards. These facts for *unions* cannot be gotten from employers; their scales do not necessarily express union experience. Unions must supply the material.

The Massachusetts Bureau of Statistics in its Labor Division collects and publishes statistics of organized labor relating to union scales of wages and hours, number and membership of unions, unemployment, strikes and lockouts, etc. Each of these will be touched upon briefly inasmuch as they probably represent the most accurate and complete data on organized labor now regularly collected by any statistical state bureau in the United States.

A report on union scales of wages and hours is regularly issued. The data are furnished entirely by unions and are published as reported, no inquiry being made as to the extent to which the union scales prevail in the various trades and

localitics. That is, minimum rates and not those actually received by union labor are published. The process of collection may be indicated by reference to the 1913 report. Returns by schedule were received from 1093 unions, or 78 per cent of those in the state. By the use of special agents 200 more were obtained, so that 92 per cent of the locals in the state were included. In tabulated form they show rates of wages by the hour, day, week, overtime (hour), and Sunday and holiday (hour); and hours of labor, by the day, week, and the period in which half-holidays are in effect, all classified for occupations and for municipalities.

Statistics on the number and membership of unions have been systematically collected and published since 1908. The collection is mainly by schedule and includes national and international unions with affiliated locals in Massachusetts, their relationship to the American Federation of Labor, the number of chartered local unions and the proportion in Massachusetts with their membership, classified for the sexes separately, by municipalities, occupations, industries, etc.

Statistics on unemployment among organized wage earners are issued quarterly. The data are collected from unions solely by schedule and are published so as to reveal the amount of unemployment by cities and occupations due to lack of work or material, unfavorable weather, strikes or lockouts, sickness, accident or old age, and other reasons, the latter specified in detail. Approximately 75 per cent of the locals are included in each quarterly report.

Statistics on strikes and lockouts have been collected by the Massachusetts Bureau since 1881. Unions and employers are scheduled on the basis of information supplied by newspapers, trade journals, etc. Besides certain preliminary data the following facts are secured from unions: the names of employers affected, conditions demanded by strikers, conditions before and granted after strikes, who ordered strikes, the occupations and numbers of strikers (the latter by sex), the dates on which strikers left and resumed work and on which strikes were

ended, as well as the methods of settlement. From employers those questions of the above which apply and the following are asked: the number of employes who struck, classified by sex; the number of non-strikers thrown out of work, classified by sex; the time lost by non-strikers; measures used by strikers to regain their positions, etc. In approximately 50 per cent of the cases the returns from the two sources are so contradictory as to necessitate the use of special agents to obtain the facts.[1] Even by this method in many cases the facts prove to be so indeterminate that the reports are published only on the basis of what *seems* to be the facts after all evidences are given their appropriate weight. These reports, therefore, appear to be summaries of reported or estimated facts concerning industrial disputes—knowledge of which is received through the press, by hearsay or by other means—having little value alone in connection with wage studies, and chiefly of interest for informational and not for functional use.[2]

Without citing further detail of the practices and experiences of American statistical bureaus in securing wage and allied data from trade unions, sufficient has been said to indicate the problems and possibilities in this approach to the study of wages. In all cases nominal and minimum rates are involved and these are reported under conditions which make it difficult, if not impossible, to apply them to unemployment data in any attempt to approximate earnings from labor service. When properly checked by scrutinizing trade agreements, nominal hours and time-rates from this source may be deter-

[1] Estimated for the writer by the Division Chief. New Jersey, placing complete reliance in newspaper clippings for initial information and *depending altogether for the facts secured on schedules from unions* alone, publishes an annual report on strikes and lockouts. If the experience of Massachusetts respecting like data is worth anything, statistics thus collected stand condemned.

[2] A detailed estimate of the value of these and like data compiled by the Bureau is not attempted here. It was made, however, by the writer during the summer of 1914 for the *United States Commission on Industrial Relations.*

mined with reasonable accuracy. Any attempt, however, to secure piece-rates on an extended scale from this source is bound to prove unsuccessful. Unemployment data from unions at best are approximations, and, of course, refer only to union labor. They serve fairly well to give a general notion of seasonal displacement of labor and of trade depression or boom but are of little value in measuring earnings or economic distress. Statistics of strikes and lockouts as collected may serve as a rough measure of the frequency of labor disturbances but not of their consequences nor of the correction which it is necessary to make for this cause when estimating wages from wage-rates.

In *summary*, we may briefly relate the statistical data extant on wages to the various concepts which this term suggests.

Comprehensive data on wages as defined above do not exist in the United States.[1] For annual reports for all manufacturing industries on classified wage-rates for short pay-periods, where conceivably wage-rates are equivalent to earnings—assuming neither over-time nor time lost—we may turn to Massachusetts, to New Jersey ("earnings" in this state), and to Ohio.[2] Studies of classified wage-rates for special industries are periodically made by the United States Bureau of Labor Statistics. In order to use nominal and minimum wage-rates as equivalent to wages it is necessary to assume that nominal conditions are actual, that figures are reported accurately, and to correct rates by figures on unemployment supplied by unions, by employers, or by employes. The reliance which can be placed in union figures on strikes and other causes of unemployment has been suggested above. The importance to be assigned to fluctuations in the employed force, as indicated by the average or actual number of employes at

[1] Nothing is said about our present national income tax statistics. The exemption allowed is so high as to omit most "wage earners," and the returns are not published in a form suitable for estimating earnings for such groups. See Falkner, R. P., "Income Tax Statistics," *Publications of the American Statistical Association*, June, 1915, pp. 521-549.

[2] Not restricted to manufacturing industries in this state.

various times in each year, depends largely upon the fluidity of labor, the ability of wage earners to find employment, and the complementary character of industries, studies of which on a significant scale have not been made. The fact of unemployment is known but it is next to impossible, except in intensive studies, to measure it by applying to those affected. The United States Census Bureau attempts to measure it from this source but the best that is secured is a rough approximation.[1] Moreover, it is chiefly among unskilled labor that unemployment is greatest, and union figures do not furnish the desired facts. Wages, therefore, in the sense in which the term is used here are not available in any other form than as estimates.

On the other hand, wage-rates for short periods, taken from employers' payrolls for manufacturing and some other industries, are reported with reasonable accuracy to a few state bureaus. In these cases, industries constitute the units, individuals and occupations being lost sight of in the grouping process. To supplement such data there are the nominal wage-rates reported by unions in which distinctions are made for occupations, industries, sexes, etc. The data are supplementary but not comparable. At least no comparisons of rates are currently published by bureaus to which both sets of facts are reported.

Earnings, in the sense of income from labor service without distinction being drawn between wages and salaries, and in contrast to property income, may roughly be approximated from the income and expenditure accounts of industrial and other businesses.[2]

[1] A question on unemployment was first included in the population schedule by the United States Census in 1880. The information secured, however, was never published. In the three succeeding censuses a similar inquiry was included, the form in 1910 being "whether out of work on April 15, 1910" and "number of weeks out of work during the year 1909."

[2] See the studies of Nearing, *op. cit.*, pp. 18-52; Streightoff, F. H., *op. cit.*, pp. 44, *passim*.

## III. A Study of Wages: Declaration of Purpose, Definition of Units, Schedule Forms

Without considering the types and sources of data on salaries and salary-rates, and without treating prices in relation to wages and wage-rates, we pass immediately, in order to illustrate the preceding treatment, to a discussion of a wage problem upon which it is intended to collect primary data. Criticism of the substance, form of tabulation, and interpretation of existing secondary data must rest with the brief sketch given above. The immediate problem, then, is to state definitely the purposes of the study which is intended to be made, to outline the plan to be followed, to define the units to be used, to formulate schedules, and to outline suggestions for the receipt and editing of returns. The precise use which will be made of the data will, of course, be determined in part by the character of the replies and can be only tentatively outlined in advance. It is intended, however, to establish certain relations and make certain comparisons between the facts reported, and the tabulations will be adjusted to these ends.

### 1. DECLARATION OF PURPOSES

The problem which has been chosen for study is the wage conditions in the textile industry in North Carolina for the year 1914. For convenience, the survey is restricted to manufacturers of cotton goods, including small wares. On the basis of information collected, schedules will be sent to 100 establishments which were found to be doing this business at some time during the year, the basis for listing establishments, separately, being that outlined in the schedules. The purpose of the questionnaire is (1) to determine the level of wage-rates for the sexes separately by age groups; (2) to measure the seasonal fluctuations in employment in relation to (a) principal product produced, (b) form of business organization; (3) to determine the total amount paid during the year in wages to employes of different sex according to an age classi-

fication; (4) to study the relation of wage-rates to (a) the form of business organization, (b) principal product produced, (c) seasonal fluctuations in number employed.

*The schedules are formulated with these purposes in mind*, and it is intended that they shall be filled in by employers without supervision other than that which is received from the instructions contained in the schedules. The study is undertaken with the assumption that it has sufficient sanction, that the filing of the returns is obligatory, that returns for individual establishments are not to be published separately, and that the results of the study will be of general social interest in which informants share equally with others. Sufficient time is to be allowed for full reports to be made and tabulations and analysis are not to be begun until satisfactory returns are received from all establishments concerned. No attempt is to be made to supplement the data collected from employers by scheduling either individual employes or unions. Complementary material may be secured from these sources but in this study it is intended to rely wholly upon returns from employers.

It must clearly be kept in mind that the discussion immediately above is illustrative of the steps which would have to be taken in the study of such a subject as wages. The facts *have been given somewhat more in detail than would have* been necessary had the purpose been merely to describe the data on wages and wage conditions in the United States. Moreover, it must be remembered that the requirement that all of the schedules must be returned is rather more severe than would be made in actual statistical work. The aim has been to set up the procedure which should be followed in an actual investigation. Of course, it is not possible entirely to do this, but the nearer it can be done, the more interest the student will have in his work and the more value he will get from it. That which is sometimes considered to be meaningless, routine clerical work may, by paralleling as nearly as can be a real problem, frequently be thought to be both necessary and vital. Great value comes from having a student

see a problem as a whole and the correlation of the different parts. By so doing the meaning of all the statistical steps through which he is led takes on new light. He is then not so much studying *method* as a problem to which method is vital in its explanation. Most mature minds desire to see some goal to their activities and reasons for the methods of study which are used. And this is as it should be, for then individuality is bound to reveal itself and the use of statistics becomes more than mere routine labor.

## 2. SCHEDULE AND EXPLANATION

### THE X. Y. COMMISSION OF NORTH CAROLINA

RALEIGH, NORTH CAROLINA

It is desired to make a study of the wages and wage conditions for the calendar year 1914 in the establishments in North Carolina which manufacture cotton goods, including small wares. All concerns in the state doing such business are included in this survey. The study is undertaken in accordance with the provisions of law, (see Chapter 673, laws 1914) and your coöperation in making it a success is respectfully solicited. Individual returns will not be published separately, and every care will be taken to hold the facts reported confidential. All employers submitting the reports called for will be furnished gratis with copies of the complete report as soon as published.

Read the whole schedule through before answering the individual questions. *Accurate* answers according to permanent records are required on *all* questions.

*Use the enclosed self-addressed and stamped envelope for returning the schedule. Schedule should be returned not later than April 30, 1915.*

THE X. Y. COMMISSION,
Raleigh, North Carolina.

I hereby affirm that the accompanying report is accurate and complete to the best of my knowledge, and is made according to the permanent records of this establishment.

| | |
|---|---|
| .......................... Name of Concern | .............................. Name of Secretary or other person making the return |
| .......................... P. O. Address | ............ Month ............ Year |

SCHEDULE TO BE USED IN THE COLLECTION OF WAGE DATA BY ESTABLISHMENTS IN THE MANUFACTURE OF COTTON GOODS, INCLUDING SMALL WARES, NORTH CAROLINA, YEAR 1914.

1. Name of Establishment.......................................

Use a separate schedule for each establishment. By an establishment is meant a plant or mill, the accounts of which are kept separately. Where separate plants are owned in common but carried on under one set of books, such separate plants are reported together as one establishment.

2. Name of Corporation, Firm, or Individual Owner...............

.......................................................

3. Location of Factory:

County .............. City or Town.................

Street and No......... P. O..........................

The location should be that of the physical plant and not of the financially controlling head.

4. Character of Business Organization (...........) (...........)
Individual Firm Partnership
(...........)
Corporation

Indicate whether individual firm, partnership, or corporation by checking thus (√) the appropriate term.

5. Frequency of Payment (............) (............). Time-
Weekly Fortnightly
or Piece-Rates (............) (............)
Time Piece

Indicate the frequency of payment, and whether time- or piece-rates *prevail* by checking thus (√) the appropriate terms.

6. Character of Industry.........................................

Indicate by giving principal product manufactured.

Please be specific respecting the principal product. The data are necessary for accurately editing the returns.

7. Number and sex of Wage Earners, both time- and piece-workers; *not* salaried employes.

Wage earners are persons receiving money or its equivalent because of manual, mechanical, or clerical labor service, paid according to a stipulated scale at frequent intervals, and under conditions which make it customary to make deductions for short periods of time lost. These should be *included*.

By salaried employes are meant persons receiving money or the

equivalent because of responsible, supervisory, or directive labor service, paid according to a stipulated scale at infrequent intervals and under conditions where it is not the custom to make deductions for short periods lost. These should be *omitted*.

| Age and Sex of Employes | A<br>Greatest Number Employed at Any Time During the Year | B<br>Least Number Employed at Any Time During the Year | C<br>Total Amount Paid in Wages During the Year |
|---|---|---|---|
| Men 18 years of age and over.. | — | — | — |
| Women 18 years of age and over | — | — | — |
| Young persons under 18 years of age | | | |
| Boys .................... | — | — | — |
| Girls .................... | — | — | — |

8. Number and sex of Wage Earners employed on the 15th of each month, 1914. If data are not obtainable for this day enter the same for the nearest representative day.

| Data to be of the 15th of the Month | Number of Wage Earners Both Time- and *Piece-workers Employed on the 15th Day* of Each Month | | | |
|---|---|---|---|---|
| | Adults 18 Years and Over | | Young Persons Under 18 Years | |
| | Males | Females | Males | Females |
| January .................. | — | — | — | — |
| February .................. | — | — | — | — |
| March .................... | — | — | — | — |
| April .................... | — | — | — | — |
| May ...................... | — | — | — | — |
| June ..................... | — | — | — | — |
| July ..................... | — | — | — | — |
| August ................... | — | — | — | — |
| September ................ | — | — | — | — |
| October .................. | — | — | — | — |
| November ................. | — | — | — | — |
| December ................. | — | — | — | — |

9. Classified Weekly Wage-rates for the Week of the Greatest Employment during the year 1914.

Do not include over-time; short-time earnings should be reduced to a full-time basis; bonuses and premiums, if any, should be included. Fines and similar deductions should be excluded.

| Specified Wage-rates Paid for the Week Ending ............ | Number of Wage Earners Both Time- and Piece-workers Receiving Specified Wage-rates Per Week | | | |
|---|---|---|---|---|
| | Adults 18 Years of Age and Over | | Young Persons Under 18 Years of Age | |
| | Males | Females | Males | Females |
| Under $3 per week | — | — | — | — |
| $3 to $3.99 per week | — | — | — | — |
| $4 to $4.99 per week | — | — | — | — |
| $5 to $5.99 per week | — | — | — | — |
| $6 to $6.99 per week | — | — | — | — |
| $7 to $7.99 per week | — | — | — | — |
| $8 to $8.99 per week | — | — | — | — |
| $9 to $9.99 per week | — | — | — | — |
| $10 to $10.99 per week | — | — | — | — |
| $11 to $11.99 per week | — | — | — | — |
| $12 to $12.99 per week | — | — | — | — |
| $13 to $13.99 per week | — | — | — | — |
| $14 to $14.99 per week | — | — | — | — |
| $15 to $15.99 per week | — | — | — | — |
| $16 to $16.99 per week | — | — | — | — |
| $17 to $17.99 per week | — | — | — | — |
| $18 to $18.99 per week | — | — | — | — |
| $19 to $19.99 per week | — | — | — | — |
| $20 to $20.99 per week | — | — | — | — |
| $21 to $21.99 per week | — | — | — | — |
| $22 to $22.99 per week | — | — | — | — |
| $23 to $23.99 per week | — | — | — | — |
| $24 to $24.99 per week | — | — | — | — |
| $25 and over per week | — | — | — | — |

REFERENCES

BERRIDGE, W. A., *Cycles of Unemployment in the United States, 1903-1922,* Houghton Mifflin, Boston, 1923, *passim.*
CHAPIN, F. STUART, *Field Work and Social Research,* Century Company, New York, 1920, Chapter VII, pp. 148-191.
KING, W. I., *Employment Hours and Earnings in Prosperity and Depression in the United States, 1920-1922,* The National Bureau of Economic Research, New York, 1923, Chapter I, pp. 9-21. (This publication also includes copies of the questionnaires used.)
NEARING, SCOTT, *Income,* Macmillan, New York, 1915, Chapter II, pp. 18-52.
RUGG, HAROLD O., *Statistical Methods Applied to Education,* Houghton Mifflin, New York, 1917, pp. 40-56.
STREIGHTOFF, F. H., *The Distribution of Incomes in the United States,* Columbia University Studies, New York, 1912, Vol. LII, No. 2.
U. S. CENSUS BUREAU, "Population Census," "Census of Manufactures."
*Business Cycles and Unemployment,* McGraw-Hill, New York, 1923, pp. 21-22; 23; 80-81; 378-387.

# CHAPTER VI

## CLASSIFICATION—TABULAR PRESENTATION

### I. Introduction

Statistical data which are to be tabulated are taken from primary or secondary sources or from both. If from primary sources, they are generally recorded on blanks used in personal interviews, on form or circular letters, or on questionnaires. In this form they are not suitable for analysis; they must be edited for consistency, accuracy, and completeness preparatory to being tabulated, averaged, and compared. If they are taken from secondary sources, some form on which to assemble them must be devised, provided the plan of arrangement in which they are found is unsuitable for that purpose. The process of orderly arranging data into columns and lines capable of being read in two dimensions is called "tabulation."

Tabulation, however, is an inclusive term. It may be discussed from three points of view: (1) the determination of the characteristics of data which are to be tabulated; (2) the manner in which they are to be classified; and (3) the form in which the classification is recorded in tables.

### II. The Characteristics of Data to be Tabulated

To place statistical data in *an orderly arrangement* presupposes a purpose. When purpose is absent, disorder is found. Before data can be orderly arranged; however, their characteristics must be determined. The questions relating to this subject are as follows:

### 1. WHAT ARE THE CHARACTERISTICS OF ANY BODY OF DATA?

The characteristics of data are their distinctive qualities or properties. Data showing the expenses of operating retail stores, for instance, vary according to volume of business done, location of the stores, their age, kinds of goods sold, types of management, etc. Farms differ in size, types of soil, ownership, productivity, position, etc.; accidents vary according to severity, nature of injury, and frequency. On the hasis of any or all of these characteristics an orderly arrangemeut can be made of such data. But other questions are immediately suggested.

### 2. IN WHAT WAY OR WAYS ARE THE CHARACTERISTICS RELATED TO EACH OTHER?

(1) They may be mutually exclusive or inclusive. For instance, a retailer's sales of suits of clothes, shoes, and umbrellas are mutually exclusive. On the other hand, his *total* sales are inclusive because they are made up of the sales of different types of merchandise. The location and fertility of farms, the age and sex of clerks, however, are mutually exclusive—they have no component parts.

(2) Some characteristics are primary while others are secondary. For instance, the total inventory value of goods on hand is secondary; the basis on which the value is taken is primary. The first depends on, or is a function of, the second.

(3) They may stand in the order of cause and effect. High wages and high operating expenses; increasing prices and increasing (dollar) volume of scales; limited production and high prices of cereals; large receipts and low prices of hogs at Chicago, etc., may be related in this way.

(4) They may be associated but not causally related as, for instance, the amount of credit sales and the volume of business done; turnover of goods and profits on sales.

(5) They may have no apparent relation to each other as, for instance, the methods of advertising specialty goods, and

the costs incurred; the methods of taking inventories and the frequency with which they are taken; the amounts of wages paid and the frequency of payment; the stature of a person and his earning capacity.

### 3. CAN DATA BE EXPRESSED IN SERIES WITH RESPECT TO TIME, SPACE, OR CONDITION?

Price differences, for instance, may be shown by days (time), by terminal markets (space), and by amount of variation or frequency of occurrence (condition).

### 4. ARE SOME CHARACTERISTICS CUMULATIVE WHILE OTHERS ARE NOT?

Amounts of sales, for instance, may be cumulated over a period of years; the customary method of paying salesmen, on the other hand, and the number of employes on the payroll of Department A in Factory B on a given pay day do not admit of such treatment.

Other peculiarities of the characteristics or properties of data will suggest themselves. What they are and the relationship between them determine the nature of the classification which is followed. But what is meant by "classification" and what does the process involve?

## III. The Nature of Classification

*Classification*, as it relates to statistics, is the process of *arranging data into sequences and groups according to their* common characteristics: of separating them into different but related parts. Some may be co-ordinate; others subordinate. It represents a process of thought—a way of analyzing a problem. The nature of the arrangement depends upon the characteristics themselves, the relations which they bear to each other, and the purpose which is to be realized in classifying them.

"Performed consciously or unconsciously, the act of classification is indispensable to and accompanies every scientific inference. A

mind is orderly or slovenly, according as it does or does not habitually and accurately classify the facts with which it comes in contact. The success of an investigation, the worth of a conclusion, are in direct proportion to the fidelity to this principle and the exhaustiveness with which the process is carried out."[1]

But what are *common* characteristics? To be "common" they must have the same properties: that is, be alike. But "likeness" is relative, not absolute. The cruder the classification, the more alike data seem to be; the finer it is, the greater the differences which are found.

The method of classifying the characteristics of statistical data can be shown by the use of examples. Certain data are available about retail stores, for instance. How may they be classified? The location, sales, expenses, inventories, purchases, and floor space are mutually exclusive categories. But each of these characteristics may be broken up into separate parts. For instance, the expenses of operation may be divided into the amounts spent for rent, wages and salaries, advertising, "busheling" (remodeling), and a number of "miscellaneous" items. The wages and salaries item is made up of amounts paid to salesmen and to proprietors, and the part paid to salesmen is composed of the amounts paid to those giving either full or part time. The compensation of full-time salesmen may be salaries or commissions, and the commissions may be fixed or fluctuating.

The employes of a factory may be similarly classified. They differ according to sex, but each sex group has its own characteristics. The males may be German or Swedish, the Swedish be native or foreign born, the foreign born be machine tenders or common laborers, and the common laborers be paid on an hourly or a daily basis.

Classification of things or the attributes of things proceeds from the general to the specific; from the most inclusive to the

[1] Cramer, Frank, *The Method of Darwin: A Study in Scientific Method,* McClurg, Chicago, 1896, p. 88.

least inclusive characteristics. Co-ordinate classes are grouped together, those which are subordinate being made subsidiary. For instance, purchases and sales are co-ordinate classes. So, also, are purchases of furnishings and of clothing, and purchases of men's and boys' clothing. On the other hand, inventories of men's suits occupy a subordinate position to inventories of men's clothing.

Whether characteristics are primary or subordinate, co-ordinate or inferior, of course, depends upon the way in which they are viewed and the purpose which is in mind in arranging them. In all cases, however, the order of thought is from the general to the specific. A logical scheme of classification is made in keeping with this general principle.

In some cases the method to be followed is established—it proceeds according to a pattern already worked out. Under such conditions, the process is automatic, clerical, routine. On the other hand, classifications are made to present, suggest, or detect relationships when they are not apparent, and when there is no guide which may be followed. Such a classification is constructive, not repetitive; creative, not clerical. To duplicate a classification is easy; to conceive one in order to test an hypothesis is difficult. It is one thing to classify the characteristics of data in keeping with instructions; it is another to determine the characteristics according to which classification should be made where no pattern is to be followed.

## IV. The Meaning of Tabulation

To tabulate data is to place them in tables—flat surfaces "with width not disproportionately small in comparison with length"—in keeping with the characteristics which have been identified and with the relations between them. The scheme involves the use of two dimensions or axes. The units in which the measurements are made generally, although not always, appear in the "caption": that is, in the vertical classes. The ways in which the measurements are presented generally, although not always, appear in the "stub"—the horizontal

classes. A tabulated datum, therefore, is found at the intersection of the vertical and the horizontal axes. It has the characteristics shown in the caption and is presented from the point of view indicated in the stub. Tabulation follows and is distinct from classification: to tabulate is to *record* data in keeping with a classification.

The tabulation form is made up of a series of "boxes," described in the captions and stub headings, into which are sorted data having the characteristics discovered through classification. The boxes or "pigeon holes" have fixed positions: they cannot be changed nor the sequences in which they are found altered without recasting the scheme of tabulation. To choose a new form, however, is not to discover new nor to discard old characteristics. They are simply *presented* in a different way.

The following statistical facts in the form presented are *not tabulated—they cannot be read in two dimensions:*

> "Employes hired during 1923: men, 536; women, 844. Withdrew, men during the year, 31; at the close of the year, 37; women, during the year, 37, at the end, 68. Men employes at beginning of 1924 from those hired during 1923, 458; those who had formerly been with the company, 51; new men, 40. Women employes at the beginning of 1924, from those hired during 1923, 739; those who had formerly been with the company, 19; and those who were new, 34."

Classification of data for purposes of tabulation, as noted above, is either automatic or experimental.

Where the form of tabulation has been determined and data are distributed according to a scheme already provided, the process is as follows:

(1) Begin with the stub. Classify the data first according to the most inclusive characteristic, and second, classify successively each subordinate part as there provided. The order of procedure, therefore, is from the general to the specific.

(2) For the most detailed characteristic in the stub, first, classify the data according to the most general characteristic

named in the caption; and second, classify successively each independent part provided in the caption. The order of classification in the caption, therefore, proceeds from the general to the specific, but in keeping with the requirements as established in the stub.

For tabulations which are not made according to fixed form, that is, for tables the purpose of which is to present, suggest, or to detect direct or associated relationships between the characteristics of data, the method is more complicated.

By a process of reasoning, trial relations between the characteristics of the data are first established. The data are then classified in keeping with these relations and distributed in a table according to caption (column) and stub (line) headings, as in (1) and (2) above. If the results which are secured are inconclusive, or of no significance—the relations which were thought to obtain not having been developed—the basis of the classification is probably without significance, although the tabulation may be correct. If this is so, it is necessary to establish other bases of classification and to follow the process of trial and error until the desired end is accomplished or proved to be impossible of realization.

In order to tabulate the data of p. 129, for instance—a form of tabulation not having been previously prepared—it is necessary to proceed as follows:

(1) Pick out the co-ordinate classes. These are as follows: men and women; years 1923 and 1924; number hired; time of withdrawals, etc.

(2) Place in the caption the classes enumerated, and in the stub the bases according to which the classes are to be distinguished or the points of view from which they are to be presented.

(3) Record in the body of the table by column and line the number of instances fulfilling the conditions named therein.

(4) Add the different parts of the co-ordinate classes. To total the columns, combine the classes in the stub; to total the lines, combine those in the caption.

The tabulated data would then appear in somewhat the form shown in Table 1.

TABLE 1

TABLE SHOWING BY SEX THE NATURE OF CHANGES IN AN EMPLOYED FORCE IN FACTORY "A," 1923 AND 1924

| Years | Change in Employed Force | Sex of Employes | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | | Men | | Women | |
| 1923 | Hired during the year 1923...... | | 1380 | | 536 | | 844 |
| | Withdrawals | | | | | | |
| | During the year.............. | 68 | | 31 | | 37 | |
| | At close .................... | 105 | | 37 | | 68 | |
| | TOTAL (deduct)............ | | 173 | | 68 | | 105 |
| | TOTAL force at end of year 1923 and beginning of 1924......... | | 1207 | | 468* | | 739 |
| 1924 | Hired during the year 1924 | | | | | | |
| | Formerly with the company... | 70 | | 51 | | 19 | |
| | New employes .............. | 74 | | 40 | | 34 | |
| | TOTAL (add) ............. | | 144 | | 91 | | 53 |
| | TOTAL at end of year.......... | | 1351 | | 559 | | 792 |

* Incorrectly given as 458.

Tables depicting the same body of data may take widely different forms. Table 1 is used only to illustrate the problem under discussion.

## V. THE ADVANTAGES OF TABULAR OVER NON-TABULAR ARRANGEMENT

Statistical data arranged in tables have definite advantages over those descriptively stated. The order in the latter case may have no logical basis; it may be according to chance or as the items were remembered or jotted down.

1. THE ORDER OF ARRANGEMENT OR THE PLAN OF PRESENTATION

When tabulations are used, some formal order is generally followed. Those most commonly used are as follows:

*(1) Arrangement According to the Size or Frequency of the Items.*

The United States Census Bureau, for instance, tabulates in a descending order the amounts of capital, values of product, etc., in manufacturing industries. The same method is followed by the Life Insurance Sales Research Bureau in tabulating by states and by districts the sales of life insurance companies. Sometimes, an ascending order is used. In either case, the method of presentation is consistent and emphatic. When the arrangement is ascending or descending, the positions of the items in the series should not be ranked by the use of consecutive numbers, as 1st, 2d, 3d, etc. The items appear in this order but the frequency or amount of the differences between them is not properly described in this manner. That this is true, in a typical case, is shown in Table 2.

TABLE 2

TABLE SHOWING THE NAMES OF INDUSTRIES AND NUMERICAL RANKING BY VALUE OF PRODUCT

(United States Census of Manufactures, 1909)

| INDUSTRIES | VALUE OF PRODUCT, 1909 | | | | |
|---|---|---|---|---|---|
| | Amount | Rank of Industry | Difference | | |
| | | | Amount | Per Cent | Rank |
| Leather, tanned, curried, and finished | $327,874,187 | 18 | | | |
| Butter, cheese, and condensed milk | 274,557,718 | 19 | $53,316,469 | 19.42 | 1 |
| Paper and wood pulp | 267,656,964 | 20 | 6,900,754 | 2.58 | 1 |
| Automobiles, including bodies and parts | 249,202,075 | 21 | 18,454,889 | 7.40 | 1 |
| Smelting and refining lead | 167,405,650 | 30 | 81,796,425 | 48.86 | 9 |

A change in rank of one, in value of product, is shown to result from an absolute difference varying from approximately seven to fifty-three and one-third million dollars, and from a relative difference ranging from 2.58 to 19.42 per cent. In one instance, a change in rank of one requires five-eighths as large an amount as is necessary in another case to occasion a change in rank of nine. In cases where it is desired to use an ascending or descending order and to indicate in a scale the positions of the different amounts, it is far better to reduce them to relative numbers, using the beginning, the last, or an average of all as a base, than to use consecutive numbers.

### *(2) Arrangement According to Time*

All data of an historical character must of necessity be presented in chronological order. The amounts or frequencies may be alike or different. This fact, however, is ignored when the time element controls. Time is continuous and unbroken, and its continuity must be preserved.

### *(3) Arrangement According to Space*

Suppose it is desired to construct a table showing by states the number of tenant farmers. The table might be arranged according to the frequency of the occurrence of this phenomenon. In this case, certain of the Southern states would, undoubtedly, occupy first place. If contiguous position were followed, the states would be listed not according to the frequency of the phenomenon, but in the order in which they occur with relation to each other. If South Carolina were listed first, Georgia and North Carolina would follow immediately. Undoubtedly, such an arrangement would be preferable to one in which neither an alphabetical, geographical, nor frequency order prevailed.

In the statistical tables of the United States Census, in-

volving geographical distribution, the order of arrangement of districts is from east to west—New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, Pacific. For the number of "Insane in Hospitals on January 1, 1910," this order is numerically roughly descending; for the percentage of population born in other divisions of the United States, the order is distinctly the reverse; and for the percentage of population under fifteen years of age it is haphazard.[1]

The relation between the phenomena described and the controlling fact in presentation—passage roughly from east to west—in these cases is not clear. It would be evident, however, in describing the distribution inland of European immigrants. Undoubtedly, arguments could be advanced for using the reverse order in describing the distribution of Asiatics in the United States. Railroad time tables invariably observe the order of contiguity. Stations are listed not alphabetically (except in the index which is *not* a table) but in the order in which they appear on the railroad line. An alphabetical order, or one according to size of city, would be of little use to one who wished to "catch a train." The point which it is sought to emphasize is that, in determining the order of data in statistical tables, account should be taken, so far as is possible, of the causal relationship or conformity which obtains between the facts tabulated and the arrangement of the data used to describe them.

### (4) *Arrangement According to a Variable Condition*

Wage-rates, income, expense of doing business, prices, interest rates, etc., are tabulated according to the frequencies with which each variation or class of variations occurs. The order is determined not by time, nor space, but by amount or degree of variation.

[1] "Insane and Feebleminded," 1910, *United States Bureau of the Census*, Washington, D. C., 1914, p. 18.

### (5) *Arrangement According to Alphabet*

No sacredness inheres in any order of arrangement except the alphabetical. But even this has its limitations. The industrial accident rate, for instance, is not necessarily highest in the "A" states, nor suicides and divorces lowest in the "U" and "W" states. It is hardly to be expected that the order of the letters in the alphabet will be of significance as a basis for distributing statistical data. And yet, this order of arrangement is frequently followed where others would be preferable. Such an arrangement is of merit as a device for identification and ready reference, but rarely otherwise.

The most emphatic parts of a statistical table are its beginning and its end. Accordingly, an ascending or descending order of arrangement is desirable in this respect. Where time, space, and frequency relations obtain, however, such an arrangement cannot be used. Moreover, no particular arrangement is best suited for all purposes. In tabulating mortality rates from tuberculosis, for instance, there would probably be an advantage in listing the districts affected according to population density, yet such an arrangement would not be suitable for all uses to which the data might be put. Nationality, mode of life, and earnings of those affected might be of more significance as a basis for grouping them. In such cases, the best order of arrangement will not be one but many. The thing that should *not* obtain is the absence of any causal or related order, and this frequently occurs when attention is not given to this detail.

Tables 3, 4, 5, and 6, showing different types of statistical data, illustrate varying orders. They should be studied to determine what, if any, considerations have controlled the arrangement. In Tables 3, 5, and 6, the occasions for using the particular orders are clear, at least for most of the classes. In Table 4 the arrangement is logical, although the basis is not so evident.

TABLE 3

NUMBER OF EMPLOYES OF RAILROADS IN SERVICE JUNE 30, 1913.*

| Class | Number |
|---|---|
| General officers | 4,398 |
| Other officers | 10,706 |
| General office clerks | 84,267 |
| Station agents | 37,721 |
| Other station men | 167,450 |
| Enginemen | 67,026 |
| etc. | etc. |

TABLE 4

RAILWAY FREIGHT CARS, NUMBER IN SERVICE, 1913.†

| Class of Car | Number |
|---|---|
| Box | 1,032,585 |
| Flat | 147,541 |
| Stock | 78,308 |
| Coal | 871,339 |
| Tank | 8,216 |
| Refrigerator | 43,389 |
| etc. | etc. |

TABLE 5

DEVELOPED WATER POWER RESOURCES, HORSE-POWER, 1900, BY DRAINAGE BASINS.‡

| North Atlantic | Horse-power |
|---|---|
| St. John River | 13,681 |
| St. Croix River | 20,500 |
| Penobscot River | 70,454 |
| Kennebec River | 63,936 |
| Androscoggin River | 123,455 |
| Presumscot River | 20,569 |
| Saco River | 25,332 |
| Merrimac River | 161,333 |
| Connecticut River | 292,899 |
| Blackstone River | 31,435 |
| etc. | etc. |

TABLE 6

NUMBER OF DEATHS IN THE UNITED STATES BY CAUSES, 1913.§

| Causes of Death | Number |
|---|---|
| Typhoid fever | 11,323 |
| Malaria | 1,565 |
| Smallpox | 125 |
| Measles | 8,108 |
| Scarlet fever | 5,498 |
| Whooping cough | 6,332 |
| Diphtheria and croup | 11,920 |
| Influenza | 7,725 |
| Other epidemic diseases | 6,382 |
| Tuberculosis of lungs | 80,812 |
| etc. | etc. |

* *Statistical Abstract of the United States*, 1914, p. 267.
† *Ibid.*, p. 266. ‡ *Ibid.*, p. 21. § *Ibid.*, p. 73.

## 2. TABULATED DATA CAN BE MORE EASILY REMEMBERED THAN THOSE WHICH ARE NOT TABULATED

Facts which are possible of association may be more readily remembered and compared when logically arranged in a table than when descriptively stated. That this is true is keenly felt when in order to make a statistical comparison one is re-

quired to read page after page of untabulated figures. The same amount of detail can generally be arranged in a table occupying only a fraction of the space and carrying much more emphasis. Respecting a certain statistical report, one critic observes as follows: "In some cases even no attempt is made at tabular presentation. Nine-tenths of the expenditure underlying statistical work that sees the light in such form has been wasted, yet some state commissions publish reams of statistics of this nature every year.* * * Thus the seventh annual report * * * contains over eighty pages * * * of closely printed statistical matter presented almost wholly in running text, without tabular arrangement." Moreover, rather *than being an aid to the understanding of a body of data, it* is deadening to have the facts contained in a table duplicated without analysis or interpretation. It is, moreover, an expensive and ineffective method of attempting to emphasize that which seems to be important.

### 3. VISUALIZATION OF GROUP RELATIONS IS FACILITATED

To group like with like into a well-arranged statistical table permits a rapid survey and a mental picture to be made of data in their different relations. When data are not tabulated, both are difficult if not impossible.

### 4. A TABULAR ARRANGEMENT MAKES IT EASY TO COMPARE DATA OF LIKE CHARACTER

To place related items in juxtaposition simplifies comparison and suggests studies which would not otherwise be thought of.

### 5. A TABULAR ARRANGEMENT FACILITATES THE SUMMATION OF ITEMS AND DETECTION OF ERRORS AND OMISSIONS

Data may be totaled when they are not in tabular form, but at considerable sacrifice of time and effort, because the items which are to be added are not placed in lines and columns.

Moreover, omissions of classes and items are not easily detected unless data are tabulated.[1]

### 6. A TABULAR ARRANGEMENT MAKES IT UNNECESSARY TO REPEAT EXPLANATORY PHRASES AND HEADINGS

The headings of lines and columns describe the items in a table. When the tabular form of presentation is not used, it is necessary, each time an item appears, to repeat the details which identify it. To do this is costly from the printer's point of view and deadening to the reader.

If it is desirable to tabulate statistical facts rather than to express them in running text—that is, to use two rather than one dimension—then it is also desirable to choose that form of tabulation which will best express the ideas which it is intended that the facts should convey.

## VI. Types of Statistical Tables

Statistical tables are of two general types: (1) *general,* and (2) *summary, derivative, or interpretive.*

General tables are detailed, their purpose being to include, so far as is possible, all of the facts which are known about the phenomena with which they deal. They are inclusive; caption and stub headings are involved and complicated, the units in which the data are expressed and the way in which they are presented serving to give a detailed account of the various properties of the data. They contain the basic "raw material," removed one or more steps from the forms upon which it is collected, and constitute the source from which summary and derivative tables may be made.

General tables are prepared when analysis is begun, their preparation constituting the first step in the process. They are sometimes little more than "working papers," to be discarded after they have served their purpose. This is almost

[1] See Table 1, p. 131.

invariably the case when summaries only are needed, and when there is no obligation felt to supply details for the purpose either of informing the public or of providing the means whereby summaries may be verified. General tables are costly to print and bulky to handle. Moreover, relatively few readers are interested in the detail which they contain. They want conclusions—"results," as they call them. Accordingly, such tables are frequently omitted from publications, separately issued, or placed in appendices.

Government bodies generally and research agencies occasionally publish such tables. In doing this they make available to others material which may be used in various ways. Interest may not lie in the particular summaries used in a statistical report; further or different analysis may be desired. In the absence of general tables, this is impossible without again collecting or assembling the data.

*But so-called "general tables" carry different amounts of* details. It is often difficult to tell whether a table is general or derivative. All tables must of necessity carry some details. Those of a summary nature, however, relate not so much to individual instances, narrow groups, and classes, as they do to totals, averages, ratios, and the like. *Summary, derivative,* or *interpretive* tables are those in which are recorded, not the detailed data which have been analyzed, but rather the results of analysis. They are brief; that is, they are in the nature of a summary. They are drawn from general tables; that is, they are derivative. They contain the results of an analysis; that is, they are interpretive. Such tables accompany the discussion of a body of data, summarizing the relations which have been found to exist among its various characteristics.

## VII. The Tabulation Form

### 1. TABLES CLASSIFIED ACCORDING TO THEIR COMPLEXITY

The form of all tables is a surface, the items being assigned to compartments in keeping with their characteristics as de-

fined in the descriptive headings in the caption and stub divisions. Simultaneously, they are read both horizontally and vertically. The greater the number of characteristics named in either caption or stub, the more complex is the arrangement of the details. On the basis of the number of divisions in captions and stubs, tables are classified as *single, double, treble,* etc.

A single table has one characteristic named in the caption and one in the stub. For instance, as in Table 7, the things named—real estate mortgages in Wisconsin—are placed in the caption, and the viewpoint from which they are presented—time—is shown in the stub.

TABLE 7

TABLE SHOWING BY YEARS THE NUMBER OF REAL ESTATE MORTGAGES IN WISCONSIN

| YEAR | NUMBER OF REAL ESTATE MORTGAGES IN WISCONSIN |
|---|---|
| Total | — |
| 1922 | — |
| 1923 | — |
| 1924 | — |
| — | — |
| — | — |
| — | — |

But real estate mortgages may be classified into two or more co-ordinate groups, as those taxable and those non-taxable, those on urban and those on rural property, etc. Similarly, each year may be divided into two or more co-ordinate parts, as January to June, inclusive, and July to December, inclusive. Tables are said to be *double* when *either* the stub or the caption contains two co-ordinate parts. Table 8 is an example of a double table, the caption being divided into two co-ordinate divisions.

TABLE 8

TABLE SHOWING BY YEARS THE NUMBER OF REAL ESTATE TAXABLE AND NON-TAXABLE MORTGAGES IN WISCONSIN

| Year | Number of Real Estate Mortgages in Wisconsin | | |
|---|---|---|---|
| | Total | Taxable | Non-taxable |
| Total | — | — | — |
| 1922 | — | — | — |
| 1923 | — | — | — |
| 1924 | — | — | — |
| — | — | — | — |
| — | — | — | — |
| — | — | — | — |

A double form may be made *treble* by providing for three co-ordinate divisions. The co-ordinate classes in Table 9 are "taxable" and "non-taxable" and "number" and "amount." The "treble" feature is due to the fact that real estate mortgages are distinguished (1) as to number and amount, (2) as to taxable or non-taxable, and (3) as to years.

TABLE 9

TABLE SHOWING BY YEARS THE NUMBER AND AMOUNT OF REAL ESTATE TAXABLE AND NON-TAXABLE MORTGAGES IN WISCONSIN

| Year | Number and Amount of Real Estate Mortgages in Wisconsin | | | | | |
|---|---|---|---|---|---|---|
| | Total | | Taxable | | Non-taxable | |
| | Number | Amount | Number | Amount | Number | Amount |
| Total | — | — | — | — | — | — |
| 1922 | — | — | — | — | — | — |
| 1923 | — | — | — | — | — | — |
| 1924 | — | — | — | — | — | — |
| — | — | — | — | — | — | — |
| — | — | — | — | — | — | — |

A *quadruple form* is secured by providing for two coordinate classes in the caption, and two in the stub; three in the caption and one in the stub; or one in the caption and three in the stub. Table 10 shows such a "quadruple" form.

TABLE 10

TABLE SHOWING BY YEARS AND BY DISTRICTS OF THE STATE THE NUMBER AND AMOUNT OF TAXABLE AND NON-TAXABLE REAL ESTATE MORTGAGES IN WISCONSIN

| YEAR | DISTRICT OF STATE | *Number and Amount of Real Estate Mortgages in Wisconsin* | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | | Taxable | | Non-taxable | |
| | | *Number* | *Amount* | *Number* | *Amount* | *Number* | *Amount* |
| Total | Total | — | — | — | — | — | — |
| | 1st | — | — | — | — | — | — |
| | 2d | — | — | — | — | — | — |
| | 3d | — | — | — | — | — | — |
| | 4th | — | — | — | — | — | — |
| | — | — | — | — | — | — | — |
| | — | — | — | — | — | — | — |
| 1922 | Total | — | — | — | — | — | — |
| | 1st | — | — | — | — | — | — |
| | 2d | — | — | — | — | — | — |
| | 3d | — | — | — | — | — | — |
| | 4th | — | — | — | — | — | — |
| | — | — | — | — | — | — | — |
| | — | — | — | — | — | — | — |
| 1923 | Total | — | — | — | — | — | — |
| | 1st | — | — | — | — | — | — |
| | 2d | — | — | — | — | — | — |
| | 3d | — | — | — | — | — | — |
| | 4th | — | — | — | — | — | — |
| | — | — | — | — | — | — | — |
| | — | — | — | — | — | — | — |
| — | — | — | — | — | — | — | — |

It will be noticed that the numbers and amounts of taxable and non-taxable mortgages are given for years and for districts. Chronology is controlling respecting time; and numerical consecutiveness, respecting space. Totals are provided for each year and for all years; for each district and for all districts. The districts are subsidiary to the years in tabular arrangement, the former being repeated under each year and the total for all years, the reason being that it is desired to compare the districts by years rather than the years by districts. Had the latter purpose prevailed, the districts would have been made primary and the years subordinate in rank. The order of arrangement respecting taxability emphasizes the direct relations between number and amount. Had the purpose been to emphasize the relation between taxable and non-taxable mortgages, the data would have been thrown into juxtaposition under the superior headings "number" and "amount."

The order of arrangement should always be that which will best develop the relations and sequences which are significant. As noted below, under *Types of Statistical Series and Corresponding Tables,*[1] the order and arrangement of data in tabulation forms should make it clear that their significance was clearly understood when the tables were planned.

Of course, more complex tables may be constructed. In fact there are no limits, except those of expense and statistical prudence, to the complexity which tabular forms may take. It is generally wise, however, to construct several tables to describe complex conditions rather than unduly to burden a single form. The amount of detail that may be grasped by the eye is limited. Too complicated tables are confusing and difficult to interpret. Judgment must be used in this instance as in all aspects of statistical studies.

## 2. TABLE STRUCTURE

While there are no hard and fast rules relating to table

[1] Pp. 157-169.

*structure, to which appeal can be made for guidance in all* cases, the following have been found helpful in getting the desired results:

*(1) Ruling and Spacing of Major and Minor Headings*

a. The amount of space assigned to major and minor headings should be in proportion to their respective importance.
b. Each subsidiary part should be given less prominence than its immediate superior. Likewise, the most subordinate heading should be assigned more space than that given to an individual item in the body of a table.
c. All forms should be set off by double lines at the top and at the bottom, the sides remaining open as they appear on the printed page. The vertical lines in the body emphasize and give distinction to the form of the table. Moreover, tables drawn in this fashion do not have a box-like appearance.
d. Major totals should be set off by double lines both horizontally and vertically. When a table is complex and divisible into two or more distinct parts, the separate portions may be set off by double lines. The complexity of form and amount of detail in each case will suggest the wisdom of modifying these general rules.

*(2) The Positions of Totals*

Totals in statistical tables were, until recently, almost nvariably placed below the detail which they summate. The Census Bureau at Washington, some years ago, began contructing tables with totals at the top, and this practice is ow quite widely followed. There is much to be said in its avor. The totals so placed are immediately before the eye nd are closely associated with the title. Almost invariably

they are of chief interest, and it is desirable to have them conspicuously placed. With totals occupying this position, totaling is upward and toward the left. The sums of totals in the lines equal the sums of totals in the columns, the check upon the accuracy showing itself in the grand total at the extreme left and upper corner of the tabular form.

*(3) Size of Tables and Suitability to the Printed Page*

The size of statistical tables is determined largely by discretion or necessity. General tables as "working papers" may be of any size desired, the only limitation being the ease with which they can be handled and the amount of detail which it is thought wise to crowd into them. If such tables are to be printed, however, the question of cost is important. Details which are thought to be necessary as a basis for thorough analysis may be considered too costly to print. Moreover, the printed page has its own limitations. It cannot be indefinitely extended. If general tables accompany the text analysis as appendices, the printed page fixes the limit of size unless folded inserts are used. If they are published separately, they should be kept within reasonable dimensions. Large pages and bulky volumes are forbidding to the average reader.

Summary, derivative, or interpretive tables, on the other hand, present no particular problems so far as size is concerned. They are generally brief and condensed and can be printed on pages of moderate dimensions. If they are too large for the width of a page, the length may be used without serious inconvenience to the reader. If too large for either dimension, readjustments of caption and stub headings—even splitting up of the table—are always possible.

From the standpoint of the reader, published tables, so far as is possible, should be included on a single page. If they run from page to page, it is necessary either to repeat in full the caption and stub designations, or to adopt some scheme of abbreviation or identification which will serve as an alternative.

*(4) The Numbering of Columns and Lines*

To number the columns and lines in general tables makes t easy to show the relationship of totals to their component arts and to verify the references to them in a text treatment. Not infrequently it is necessary in text analyses, when referring o items in detailed tables, to employ awkward descriptive hrases where it would be easy, by citing line and column umbers, unmistakably to fix their position. One often hesitates to verify references to items because of the time involved i identifying them. The costs and inconvenience of numberig both columns and lines are so small, while the value is so naterial, that it seems desirable to adopt both practices in all ables in which the amount of detail is large or the form of ie tabular arrangement at all complex.

As an alternative to using guide or margin numbers—line umbers—some of the United States statistical publications range lines into groups of five. This breaks up the detail id relieves the monotony of an elaborate table, thus making easier to follow, but it does not solve the difficulties in text ialysis of referring to the details in general tables and of owing the columns which are summarized into totals. olumn numbers, moreover, often help to interpret the relans between the items in a detailed table. These are not ways self-evident even to those experienced in statistical idy.

## VIII. The Contents of Tables

The contents of a table, obviously, have to do with the rpose which it is intended to serve. If it constitutes a form record only, the data will be detailed; if it serves as a type analysis, they will be abridged and summarized. Whatever purpose, the contents should be determined in keeping with following rules:

(1) They should relate solely to the purpose in mind. Extraneous materials should not be included: they detract m those which are of interest. Moreover, the relations of

those which are included to the purpose to be accomplished by including them should be evident. Every table should be easily understood, and the relation of each part to the whole and to the other parts be apparent.

(2) The items should be accurately distributed and the totals correctly summated.

Totals are but the functions of the items which compose them. They are generally no more accurate than the items unless errors compensate each other. This condition rarely occurs. As to whether it does in a particular case may be determined only by a study of the units in which the measurements are made; the purpose, plan, and motive governing their collection; the interpretation assigned them, etc.—topics described at length above. The discovery of unexplainable errors in a table itself raises a presumption against the accuracy of all of the preceding stages through which data have been carried. Moreover, unless its nature is known and can be allowed for, it makes doubtful the use of subsequent tables into which the error may have been carried. A known error can be corrected; one which is unknown is compromising at every turn. Totals should be made to cross-check accurately, account being taken of the possibility that compensating errors may appear in both lines and columns and still the cross-check agree. A cross-check is not a complete guaranty that inaccuracies do not exist within the body of a table.

(3) Summary, derivative, or interpretive tables, so far as possible, should carry references to (a) the meanings of the terms which are employed; (b) the pages from which the summaries are taken, and the table, line, and column numbers involved; and (c) the scope of the data summarized or averaged.

(4) Statements of the peculiar meaning and limitations of statistical tables should closely accompany the tables themselves, be conspicuously placed and clearly stated.

No one is as well prepared to know the limitations of data, at each stage of collection and tabulation, as he who pre-

pares them, and, in justice to all, they should be clearly stated. The *place for an appraisement to appear is where no one can* overlook it.

(5) "Miscellaneous," "not stated," and "unclassified items" in statistical tables should be kept at a minimum.

In case such classes are numerous, it is a wise precaution against misunderstanding and a valuable aid in interpretation to add an explanatory note showing in a general way their contents. Normally, such notes do not immediately accompany tabular forms, with the result that they are overlooked.

(6) Tables should be arranged, so far as possible, so that items will appear in each compartment named in the caption and the stub headings.

(7) Averages, ratios, etc., should not be made a conspicuous part of general tables. They should be reserved for those which are of a summary, derivative, or interpretive nature. The two types of tables, of course, are not always distinct. In some cases, particularly in brief studies, they shade imperceptibly into each other, the same table serving both for purposes of record and of summary. In all but the briefest studies, however, differentiation can be made and is desirable. It is far better to have a complete statement of the limitations of the data, adequate definitions of the units and reasons for the combinations which are made of them given in general tables, than it is to dispense with them and have the tables filled with averages and percentages. It is the function of *the statistician to* make *statistical* data as comprehensive and full of meaning as they can be made. It is not his purpose, in connection with general tables, to analyze them: this function is reserved for summary tables. Much time, effort, and money are wasted in crowding into general tables an elaborate network of percentages, averages, and the like.

## IX. Titles for Statistical Tables

The title of a statistical table should be a brief epitome of its contents. The most important categories should be

specifically named but no attempt made to include all of the different characteristics. It is not the purpose of a title completely to summarize the contents of tables. It should be short, clearly phrased, well punctuated, and impossible of double meaning. Titles are generally faulty because of omissions, improper phrasings, and inverted order. Normally, the things enumerated in the title should follow the order of the superior and subsidiary headings. For instance, if a table has to do with wage-rates, classified on hourly, daily, and weekly bases, and these are presented by occupations and by districts, or by the nationalities of those occupied, then this order should be followed in the title. To invert the order is confusing and may be misleading.

Illustrations of faulty titles, omissions of column headings, and other details to be guarded against in tabulations might be cited at length but the following will suffice for this purpose. The reader should always be on the lookout for errors and bad form in statistical presentation. In this way he is able to improve his own methods and to benefit by the mistakes of others.

In Table 11, co-ordinate classes in the caption are not given equal prominence. These classes are "Fatal" and "Non-Fatal." Accordingly, they should be made to appear of equal importance, the detail of non-fatal accidents being reduced to a subordinate position.

In Table 12, there are three co-ordinate classes, but this fact is not apparent from the arrangement of the table. Moreover, "Lacerations or Abrasions" are placed as subordinate to "Fingers Cut Off," and "Hand Cut Off" is placed between the details of "Fingers Cut Off" and "Total Fingers Cut Off." This arrangement is wrong. Moreover, the total should indicate the number of "individual" accidents, because, for instance, the loss of four fingers is called four accidents.

*Omissions in Column Headings*

TABLE 11

THE CAUSES OF ACCIDENTS RESULTING IN INFECTION

| | TOTAL | FATAL | AMPUTATIONS | INFECTED CUTS AND PUNCTURES | INFECTED BRUISES | INFECTED BURNS | INFECTED EYES |
|---|---|---|---|---|---|---|---|
| Causes of accidents | 721 | 5 | 4 | 511 | 102 | 53 | 46 |
| Nails in floor | 32 | 1 | — | 31 | — | — | — |
| — | — | — | — | — | — | — | — |

The above table should have been constructed thus:

| CAUSES OF ACCIDENTS | TOTAL | FATAL TOTAL | NON-FATAL | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | TOTAL | AMPUTATIONS | INFECTED CUTS, ETC. | INFECTED BRUISES | INFECTED BURNS | INFECTED EYES |
| Total...... | 721 | 5 | 716 | 4 | 511 | 102 | 53 | 46 |

*Misplaced and Confusing Headings and Totals*

TABLE 12

JOINTER ACCIDENTS REPORTED, BY NATURE OF DISABILITY

| GUARDED OR UNGUARDED MACHINES | ALL ACCIDENTS | TOTAL FINGERS CUT OFF | HAND CUT OFF | FINGERS CUT OFF | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Four fingers | Three fingers | Two fingers | One finger | LACERATIONS OR ABRASIONS |
| All accidents | 77 | 71 | 1 | 4 | 2 | 11 | 27 | 32 |
| — | — | — | — | — | — | — | — | — |

This table should have been arranged thus:

| CAUSES OF ACCIDENTS | TOTAL INDIVIDUAL ACCIDENTS | HAND CUT OFF | LACERATIONS | FINGERS CUT OFF | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Total | Four | Three | Two | One |
| Total...... | 77 | 1 | 32 | 71 | 4 | 2 | 11 | 27 |
| — | — | — | — | — | — | — | — | — |

*Faulty Rulings and Misplaced Column Headings*

TABLE 13

ACCIDENTS CAUSED BY FALLS OF WORKMEN—BY CAUSE AND DISABILITY

| CAUSES OF ACCIDENTS | TOTALS | PER CENT DISTRIBUTIONS | FATAL | LOSS OF FINGERS | INTERNAL INJURIES | FRACTURES | SPRAINS | LACERATIONS | BRUISES | BURNS | INJURED EYES |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total—all Causes ... | 1387 | 100.0 | 48 | 2 | 30 | 425 | 384 | 110 | 346 | 41 | 1 |
| Falls down | 52 | 3.7 | — | — | — | 19 | 15 | 5 | 13 | — | — |
| — | — | — | — | — | — | — | — | — | — | — | — |

*The total columns should have appeared thus:*

| CAUSES OF ACCIDENTS | TOTAL | |
|---|---|---|
| | Number | Per cent Distribution |
| Total......... | 1387 | 100.00 |

## X. THE MECHANICS OF TABULATION

Before the actual process of tabulation is begun, it is generally necessary to prepare data for tabulation. It is almost never possible immediately to transfer them from schedules or other primary records onto tabular forms. Data must first be *edited*. Errors must be corrected, omitted items filled in, conflicting statements harmonized, and consistency secured. This does not mean that the data have to be "cooked." Not at all. They are simply reduced to a comparable basis so that they may be combined into groups and classes.

After data are edited, they are frequently "coded," individual numbers or letters being assigned each separate group and characteristic. By the use of such codes, long descriptions

and involved class distinctions are abbreviated, the numbers and letters standing in place of the original terms and serving to identify them.

The coded data are then transcribed onto tabulating cards. These may be designed for hand or for machine use, but, howsoever employed, they have among other the following characteristics:

(1) A given space on a card is always reserved for a particular entry.

(2) Each separate card or a series of them has to do with a single report, reporting agency, or condition.

The cards for hand tabulation may be designed at will. Those commonly used are either three inches by five inches, or five inches by eight inches, the surfaces being divided into as many separate divisions as are necessary to include the data to be tabulated. Cards of larger size are sometimes used, but the smaller sizes permit of greater accuracy and speed in sorting. It is difficult to sort large cards for items appearing in the central blocks. The arrangement of the parts may follow any order, but that which is most logical should be chosen. The logical order is generally the same as that followed in the questionnaire, although it may be desirable at times, in order to group together related items, to choose a different arrangement.

In a recent study[1] six hand and one machine tabulation cards were necessary to record all of the data available. The plan of arrangement of the detail on the cards did not follow that used in the schedules. The basic facts were placed on Card 1, the others carrying less significant detail. The form of Card 1 is shown in Figure 2.[2]

[1] *Costs, Merchandising Practices, Advertising and Sales in the Retail Distribution of Clothing*, Bureau of Business Research, Northwestern University, Prentice-Hall, Inc., New York, 1921.

[2] The respective letters and numbers refer to subject, page and inquiry. For instance, the third block, Pop-C-2,2(1), has reference to population of city, page 2, inquiry 2(1).

FIGURE 2

HAND TABULATION CARD

| Scn | Bus-£,1 / Pop-C-2,2 (1) | C-2,2 (3) | Sc-2,2 (4) / I-2,2 (6) | Fl-3,3 (8) | Change Window 8,3 (12) | | 1. |
|---|---|---|---|---|---|---|---|
| 1914 | Sales - 4, T. | Mo.Act-5,(2)B | Fixt.-7,6 (7) | Av.Stock 7,(9) T | Sal-11,2T | Bush-15,5T | P.B.-3,4(2) |
| 1918 | | | | | | | |
| 1919 | | | | | | | |
| 1914 | Return-4,5Q | Purch-6,T | Deliv-7,6(8) | Tot.Exp-10A | Adv-13,T | Tax-15,III | Bld. 10,1(1) |
| 1918 | | | | | | | |
| 1919 | | | | | | | |
| 1914 | Charge-5,(2)A | Disc. 6, Q | Inven-7,(9)T | Rent-10,1(2) | Gen.Exp.14.T | Cap.Exp.15, IVT | |
| 1918 | | | | | | | |
| 1919 | | | | | | | |

Hand cards may be used to advantage when

(1) the number of instances to be tabulated is comparatively small.

(2) the items are large quantities and when it is necessary *to record exact amounts.*

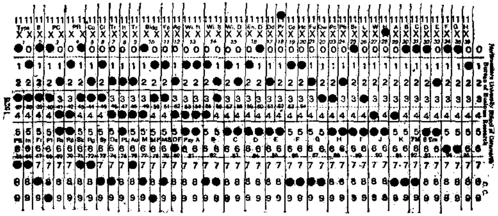
(3) it is desirable or necessary to compute on the cards ratios or averages.

Tabulation cards suitable for machine use may also be employed. The best known are the "Hollerith," furnished by the Tabulating Machine Company. Both machine and hand cards are alike in principle—a given position always having reference to the same fact, but not the same phase in which it is encountered. The cards are provided in blank—the face being covered with a series of numbers in lines and columns —or they are specially prepared to suit a given code system. In either case, they are used in essentially the same manner as are the hand cards: that is, sorted into groups or classes according to the code designations in keeping with a scheme of

tabulation. On machine cards, the presence of an item or ar amount is shown by perforation; on the hand card, by a symbo or the fact itself. Machine cards are sorted and totaled by electrical contacts; hand cards, by hand methods.

FIGURE 3

MACHINE TABULATION CARD



A *facsimile* of the machine card used in the study referred to above is given in Figure 3.[1] Machine tabulation cards may be advantageously used when

(1) *the number of instances to be tabulated is large.*

[1] The details about the store to which this card refers are as follows: This store handles clothing and furnishings only; is not a department store; is located in a city with population between 20,000 and 40,000 in 1920; has a trading radius with population of 120,000 to 200,000; is on a corner; on a street car line; not at a transfer point; is on an interurban line but not at a transfer point; is constructed of brick, and is fireproof; the building is between 25 and 40 years of age; the total linear feet of window display space is between 20 and 30 feet; it has vestibule windows without islands; the depth of the windows is 7 feet; the depth of the entrance is between 6 and 9 feet; it has no double-deck windows; uses clothing and hat cabinets and furnishing goods units; has been in business between 10 and 18 years, all of this time in the same city, and 6 to 10 years in this building; it has no branches; the length of floor space above the basement is between 90 and 100 feet; the width of the floor space is between 30 and 40 feet; the height of the lowest floor above the basement is 15 feet; is located entirely on the first floor with between 3000 and 4000 square feet of floor space; the basement area is between 1000 and 2000 square feet; has no mezzanine floor or balcony; the total area of floor space (including stock rooms) is between 4000 and 6000 square feet.

(2) the amounts can be arranged into groups, and only the group designation indicated.
(3) the class items are mutually exclusive, and can be indicated by symbols.
(4) all of the data can be placed on one card.
(5) tabulations of the same type are recurrent.

Some of the advantages of using the "card" system in tabulation are as follows: (1) Any combination of characteristics is easily made; (2) each characteristic or amount is always assigned the same position on the card; (3) the cards are always available for tabulation.

After data have been coded and transcribed onto suitable cards, they are then sorted according to the characteristics which it is desired to tabulate. The accuracy with which punched cards are sorted may be checked by holding the cards up to the light and noting whether it passes through the respective holes for the different items. Any obstruction of the light automatically registers an error in sorting. The accuracy of the sorting, when done by hand, may be checked by turning through the cards and scrutinizing each of them for errors. In order that this may be done conveniently, the cards must be relatively small and the edges accurately cut. Punched

(Note 1 continued)

The store sells boys' and children's clothing, men's furnishings, boys' and children's furnishings, men's hats and caps, boys' and children's hats and caps, work clothing; it does not sell men's and boys' shoes, men's fur goods, luggage, women's wear, nor women's shoes; its sales of work clothing include overalls, union-alls, denims, cotton suits, jackets, work shirts; from 2 to 10 per cent of its sales are of palm beach; it takes its inventory at depreciated values; does not add freight and other charges to inventory value; does not keep a perpetual inventory record; does not keep a record of prices nor sizes; uses sales books, a cash register, no patented system of books of account; does not keep a daily record of profits; prepares a monthly profit and loss statement; has its accounts audited annually by outside accountants; charges to personal account all merchandise taken out for personal use; charges 10 to 20 per cent depreciation on fixtures; pays its buyer, regular- and extra-salesmen, bookkeeper, window trimmer, advertising man, and bushelmen straight salaries; does not use P.M.'s; sells goods to its employes at cost; had sales during 1919 between $140,000 and $180,000.

cards may be employed to advantage even where electrical machines for sorting or counting are not available. Cards are sorted first into the more comprehensive groups and subsequently into the sub-groups provided for in the scheme of tabulation.

After the cards have been sorted, the next process is to count or add the frequency of the occurrence of each item. This may be done in connection with the tabular form when direct transcription is made from the schedule or original sheet to the table. When large aggregates must be summated before tabular entry can be made the process is not easy without first listing the facts. The use of adding machines for this purpose is imperative. It is best to use a listing machine and to retain the sheets for future reference. When comparisons are to be made, the items on the listing paper may be used in computing percentages, averages, etc., for making new combinations, and for cross-checking.

It is frequently necessary to arrange data into groups and to express the occurrence of each item in a frequency table in the manner described immediately below. In so doing, the individual instance *per se* is lost sight of. This need is particularly true respecting data on wages, sales, ages, etc.—cases in which it would be difficult, if not impossible, to take account of the precise measure of each individual instance. The listing or tallying may be done by arranging on the left-hand margin of a sheet of paper the groups into which the individual items are to be placed, and by tallying off opposite each individual group the number of instances occurring. This method has the disadvantage of making impossible any check on the accuracy of the work. An alternative method is to transcribe the data to be grouped onto small cards and to arrange them into groups, thus allowing each group to be checked by rapidly running through the cards. This method requires that the data be copied, thus allowing error to enter from this source. Whichever method is followed, the accuracy of the listing should be thoroughly verified.

## XI. Types of Statistical Series and Corresponding Tables

Statistical series[1] are of three types: (1) historical, (2) spatial, and (3) condition. Corresponding to each of these types are the tables in which they are tabulated. Accordingly, there are tables showing data with respect (1) to their time relations, (2) their space relations, and (3) with respect to the frequency of occurrence of things or the attributes of things at a given time and space. These different series with their corresponding tables require brief consideration.

The controlling factor in tabulations which express historical series is, of course, chronology. Normally, the arrangement is simple and easily comprehended. All of the facts, no matter how diverse in frequency or divergent in type, are tabulated according to time. Only when time is significant, however, should chronology dominate the arrangement of statistical detail. In cases where it is incidental it should be reduced to a subsidiary position. The degree of prominence to be given to it depends in each case upon the purpose of the table.

In tabulating space series, the controlling factor in presentation is place or location. Variation is seen geographically. Chronology has no significance since measurements varying in relation to space are taken as of a given time. Table 14 represents such a series. The data in this table refer to a given period of time, and show the methods of wage payment and the rates of wages in different municipalties. That is, the table presents statistical series viewed geographically, an alphabetical arrangement being followed.

Of course, a contiguous rather than an alphabetical arrangement of the cities might have been followed. Such an order would be preferable to the one followed if the wage-rates were in any way related to the location of the municipalities. More-

[1] A "series," as used statistically, may be defined as things or attributes of things arranged according to some logical order.

TABLE 14

TABLE SHOWING UNION SCALES OF WAGES FOR PLUMBERS ON OCTOBER 1, 1913, BY MUNICIPALITIES (*LABOR BULLETIN* No. 97, MASS. BUREAU OF STATISTICS, p. 39, BOSTON, MASS.)

| MUNICIPALITIES | RATES OF WAGES | | | | |
|---|---|---|---|---|---|
| | Hour | Day | Week | Over-time (hour) | Sundays and Holidays (hour) |
| Attleborough ...... | \$0.40⅝ | \$3.25 | \$19.50 | \$0.81¼ | \$0.81¼ |
| Beverly .......... | .60 | 4.80 | 26.40 | .90 | 1.20 |
| Boston ........... | .62½ | 5.00 | 27.50 | 1.25 | 1.25 |
| — | — | — | — | — | — |

over, the space units might have been listed according to size, but only on condition that there were some relation between the details and the size of the cities. Before any arrangement is chosen, the relations which it is desired to emphasize should be clearly determined. Tabulation is rarely the first step in analysis, frequently it is the last step, the early ones having been taken in deciding upon the form to be used. A large part of the exposition necessary to make plain what is intended to be shown can be obviated if a table on its face unmistakably reveals its purpose. There is nearly always a *best* form, and it is the peculiar function of the person using statistics to discover it. After all, a table is only a form on which are recorded relations and sequences.

Condition series constitute a third type of statistical series, the corresponding tables being known as "frequency tables." Variation in size and amount characterizes statistical measurements of things and their attributes. Uniformity rarely obtains. The different measurements of natural phenomena are distributed about a norm or common measurement when a large number of instances are taken, or when sufficient samples are chosen purely at random. If, for instance, one were to measure the lengths of a number of leaves, chosen at random from a particular tree, the different measurements would vary,

although a most common or characteristic length would be found. From this, other measurements would deviate, some being longer and some shorter. If a large number were taken and pure chance governed their selection, the number of those having lengths greater than the characteristic or common measure would tend to be equal to those having lengths shorter than the standard as determined. A tendency toward uniformity of distribution in excess and in defect of a common measure characterizes all natural phenomena.

A similar regularity of distribution results from measuring the same thing a number of times. Each measurement is influenced by the "measuring stick" and by the way in which it is used. With successive trials, however, the errors due both to physical and human causes will tend to be eliminated or corrected, and a common or characteristic result be secured. With pure chance operating, the deviations or "errors" will be distributed in excess and in defect of the "true" measurement in a systematic and regular order, those in excess tending to equal those in defect.

In the measurements of economic phenomena, a like tendency for variations to be systematically distributed about a norm is observed. Wage-rates vary within narrow margins for the same type of labor for a given district, and between districts the differences are not large. For a given occupation, a norm is established. Wage-rates above and below this standard are exceptional both as to the amounts and the number of individuals receiving them. The foot frontage value on a certain residence city street varies only within a narrow margin, the amount of deviation from the extremes being relatively small and the frequencies relatively few. Down-town business blocks have a characteristic height. Few will be higher than twenty stories, and few less than three stories high. Most American freight cars have a capacity of from thirty to fifty tons; very few now in use for freight services have a capacity of less than fifteen tons, while few are built with a capacity beyond one hundred tons. The ruling interest rates

on real estate mortgages range from 5 to 6 per cent. Some loans are made at less than 3 per cent, and a few others at more than 10 per cent. The most characteristic rate is probably 5 per cent. A norm in such cases tends to be established, but it does not obtain in the same rigorous fashion in economic as it does in natural phenomena.

In tabulating such variable phenomena, frequency tables are used. Such tables are constructed by listing singly or in groups and according to ascending order the units in which a phenomenon or condition is measured, and by arranging opposite them the corresponding frequencies with which they occur. Tables 15 and 16 will serve as illustrations.

TABLE 15

FREQUENCY TABLE SHOWING CLASSIFIED WEEKLY WAGES FOR EMPLOYES IN ALL MANUFACTURING INDUSTRIES IN MASSACHUSETTS, 1912

(*27th Annual Report*, Statistics of Manufactures of Massachusetts, 1912, p. xxii, Boston, Mass.)

| WAGE GROUPS | NUMBER AND PER CENT OF EMPLOYES RECEIVING SPECIFIED AMOUNTS | |
|---|---|---|
| | Number | Per cent |
| Total | 681,383 | 100.0 |
| * Under $3 per week | 2,266 | 0.3 |
| * $3 but under $4 | 5,792 | 0.9 |
| $4 but under $5 | 16,909 | 2.5 |
| $5 but under $6 | 34,070 | 5.0 |
| $6 but under $7 | 52,604 | 7.7 |
| $7 but under $8 | 63,879 | 9.4 |
| $8 but under $9 | 68,787 | 10.1 |
| $9 but under $10 | 75,006 | 11.0 |
| * $10 but under $12 | 103,160 | 15.1 |
| * $12 but under $15 | 107,677 | 15.8 |
| * $15 but under $20 | 104,585 | 15.3 |
| $20 but under $25 | 32,536 | 4.8 |
| * $25 and over | 14,112 | 2.1 |

* Note the changing widths of the groups and the treatment of the residuum.

## TABLE 16

FREQUENCY TABLE SHOWING THE NUMBER OF DEATHS FROM ALL CAUSES

Registration Area, United States, 1912 (*Mortality Statistics*, 1912, p. 11, Washington, D. C., 1913)

| AGE OF DECEDENT | NUMBER | | |
|---|---|---|---|
| | Total | Male | Female |
| All ages | *838,251* | *459,112* | *379,139* |
| * Under 1 year | 147,455 | 82,834 | 64,621 |
| * 1 year | 29,713 | 15,748 | 13,965 |
| * 2 years | 13,189 | 6,889 | 6,300 |
| * 3 years | 8,240 | 4,392 | 3,848 |
| * 4 years | 6,042 | 3,178 | 2,864 |
| † Under 5 years | *204,639* | *113,041* | *91,598* |
| 5– 9 years | 17,274 | 9,149 | 8,125 |
| 10–14 years | 11,436 | 6,008 | 5,428 |
| 15–19 years | 20,343 | 10,525 | 9,818 |
| 20-24 years | 30,997 | 16,696 | 14,301 |
| 25-29 years | 33,762 | 18,495 | 15,267 |
| 30-34 years | 33,743 | 18,929 | 14,814 |
| 35-39 years | 37,916 | 21,850 | 16,066 |
| 40–44 years | 37,885 | 22,337 | 15,548 |
| 45–49 years | 39,624 | 23,638 | 15,986 |
| 50–54 years | 45,496 | 26,995 | 18,501 |
| 55–59 years | 45,732 | 26,451 | 19,281 |
| 60–64 years | 51,097 | 28,637 | 22,460 |
| 65–69 years | 55,492 | 30,045 | 25,447 |
| 70–74 years | 55,650 | 29,219 | 26,431 |
| 75–79 years | 50,772 | 25,808 | 24,964 |
| 80–84 years | 36,678 | 17,689 | 18,989 |
| 85–89 years | 19,559 | 9,027 | 10,532 |
| 90–94 years | 7,082 | 2,997 | 4,085 |
| 95–99 years | 1,493 | 620 | 873 |
| ‡ 100 years and over | 458 | 169 | 289 |
| ‡ Unknown | 1,123 | 787 | 336 |

* Note the lower groups.
† Note the summary of lower groups.
‡ Note the residuum and the "Unknown."

When units of measurement are grouped, accuracy of detail may or may not be sacrificed. If a series is *discrete* any grouping serves to disguise the truth; if a series is *continuous*, it may aid in revealing it.

By *continuous* series are meant those in which measurements are only approximations, within the limits set up, to an absolute but indeterminate value. By *discrete* or broken series, on the other hand, are meant measurements which are determined by the nature of the units themselves. In continuous series, measurement is dependent upon the accuracy with which approximations are made. In discrete series, measurements are determined simply by the nature of the units.

The following series of measurements are *discrete:* the number of rows of kernels on ears of corn; the number of pages in books; the number of letters in words; prices at which books are sold; the wage- and salary-rates paid to employes; the number of "parts" in automobiles.

On the other hand, the following series are *continuous:* the weights of bushels of corn, wheat, etc.; the weights of hogs received at Chicago on a given day; the square feet of floor space used in grocery stores; the ages of workingmen; the length of time it takes different men at the same time or place, or the same man at different times or places, to put threads on a bolt.

Both time and space units, as such, are always continuous, but the measurements of phenomena in time and space may be continuous or discrete. The number of books sold per year, for instance, may be determined. The facts are discrete. The time in which they are sold, known as a "year," however, is continuous. Its limits are arbitrarily determined. On the other hand, not only may the unit of time but also the measurement which is expressed in time be continuous. Such measurements as temperatures at hourly intervals constitute an example. Heat and cold exist not as absolute but only as relative conditions.

Similar observations also apply to space measurements.

Space itself is continuous, but the measurements of phenomena in space refer to things or their attributes which are continuous or discrete. Numbers of employes by departments are discrete; ages, for the same population, are continuous. Again, the number of tractors per farm is discrete; the number of acres per farm is continuous.

The distinction between discrete and continuous measurements so far as tabulation is concerned, however, is chiefly of interest where neither time nor space, but variation at a time or within a space is involved.

The example of a discrete series in Table 17, showing the number of real estate mortgages in Wisconsin in 1904, classified by rates of interest, illustrates the relations between frequencies and units of measurement, and the effect which different widths of groups have upon the frequencies.

A study of the distribution shows that the frequencies in groups beginning with the half per cents and extending to but not including the even per cents are conspicuously less than in those beginning with the even per cents and extending to but not including the half per cents. The numbers in the former groups show not only a greater concentration on the even than on the half per cent units, but also a greater concentration on the half per cent than on any other fractional units. The frequencies are determined by the units in which interest rates are commonly expressed, and there is no reason why an equal distribution throughout the widths of the groups should be expected. There is nothing in the nature of the measurements which requires the units to be continuous and infinitesimally small.

As the groups stand in column (a), the piling up of the frequencies on the lower side is evident in every case. If they are widened, as in column (b), the distribution is still of the same general character; but the relative degree of concentration on the half per cent and other fractional parts cannot be determined. Column (b) is distinctly less suggestive for the separate groups, but much more so for the complete range than

is column (a). In the distribution in column (c)—one per cent groups, as 3½ but less than 4½ per cent, etc.—the even per cents appear in the middle of the groups, the emphasis assigned to them being theoretically distributed over the whole group. This theoretical dispersion does not, however, fit the case; *the concentration is still on the even per cents*, and any attempt to distribute it evenly over the whole group conflicts with the facts as shown in column (a). For purposes of analysis, it is often desirable to place the limits of the groups as in column (c), but it is always necessary to remember the actual as distinct from the theoretical distribution.

TABLE 17

Frequency Table Showing the Number of Real Estate Mortgages in Wisconsin, 1904, Classified by Rates of Interest

| Rates of Interest | Number of Real Estate Mortgages | | |
|---|---|---|---|
| Total | 28,961 (a) | 28,961 (b) | 28,961 (c) |
| Under 3% | 35 | 35 | 35 |
| 3 and less than 3½% | 133 | 164 | 133 |
| 3½ and less than 4% | 31 | | 1,309 |
| 4 and less than 4½% | 1,278 | 1,785 | |
| 4½ and less than 5% | 507 | | 10,769 |
| 5 and less than 5½% | 10,262 | 10,878 | |
| 5½ and less than 6% | 616 | | 10,004 |
| 6 and less than 6½% | 9,388 | 9,621 | |
| 6½ and less than 7% | 233 | | 4,531 |
| 7 and less than 7½% | 4,298 | 4,327 | |
| 7½ and less than 8% | 29 | | 1,639 |
| 8 and less than 8½% | 1,610 | 1,615 | |
| 8½% | 5 | | 60 |
| 9% | 55 | 56 | |
| 9½% | 1 | | 478 |
| 10% | 477 | 477 | |
| 12% | 2 | 2 | 2 |
| 16% | 1 | 1 | 1 |

## TABLE 18

FREQUENCY TABLE SHOWING DISTRIBUTION OF THE LENGTHS OF LOBSTERS *

| LENGTHS IN INCHES | (Frequency) (a) | ½ INCH GROUP (Frequency) (b) | ¾ INCH GROUP (Frequency) (c) | 1 INCH GROUP (Frequency) (d) | 1 INCH GROUP (Frequency) (e) |
|---|---|---|---|---|---|
| 8 | 6 | 8 | 11 | 14 | 6 |
| 8¼ | 2 | | | | |
| 8½ | 3 | 6 | | | 151 |
| 8¾ | 3 | | 181 | | |
| 9 | 143 | 178 | | 474 | |
| 9¼ | 35 | | | | |
| 9½ | 241 | 296 | 810 | | 845 |
| 9¾ | 55 | | | | |
| 10 | 514 | 575 | | 1152 | |
| 10¼ | 61 | | 638 | | |
| 10½ | 582 | 577 | | | 1206 |
| 10¾ | 45 | | | | |
| 11 | 568 | 611 | 918 | 929 | |
| 11¼ | 43 | | | | |
| 11½ | 307 | 318 | | | 775 |
| 11¾ | 11 | | 433 | | |
| 12 | 414 | 422 | | 590 | |
| 12¼ | 8 | | | | |
| 12½ | 156 | 168 | 489 | | 497 |
| 12¾ | 12 | | | | |
| 13 | 321 | 326 | | 474 | |
| 13¼ | 5 | | 153 | | |
| 13½ | 146 | 148 | | | 579 |
| 13¾ | 2 | | | | |
| 14 | 426 | 426 | 516 | 516 | |
| 14¼ | | | | | |
| 14½ | 90 | 90 | | | 370 |
| 14¾ | | | 281 | | |
| 15 | 280 | 281 | | 329 | |
| 15¼ | 1 | | | | |
| 15½ | 45 | 48 | 151 | | 152 |
| 15¾ | 3 | | | | |
| 16 | 103 | 104 | | 117 | |
| 16¼ | 1 | | 14 | | |
| 16½ | 13 | 13 | | | 44 |
| 16¾ | | | | | |
| 17 | 30 | 30 | 33 | 33 | |
| 17¼ | | | | | |
| 17½ | 3 | 3 | | | 10 |
| 17¾ | | | 7 | | |
| 18 | 7 | 7 | | 7 | |
| 18¼ | | | | | |
| 18½ | | | 4 | | 4 |
| 18¾ | | | | | |
| 19 | 4 | 4 | | 4 | |
| 20 | | | | | |

* The measurements in column (a) are taken from the *American Statistical Association Publications*, Vol. 7, p. 60. The original data are in a monograph by Dr. Francis H. Herrick on "The American Lobster in the United States," *Fish Commission Bulletin* for 1895.

In contrast with series such as that given in Table 17 which is discrete both as to the unit (interest rate) and the measurement (the number) are those which are continuous in one or in both respects. In Table 18, showing the number (the measurement) of lobsters of different lengths (the unit being length to the nearest quarter of an inch), the unit is continuous and the measurement discrete. In classifying these *crustacea, the measurements are first distinguished* by quarter inch differences. When this is done, the frequencies as in column (a) are unevenly distributed for lengths approximately equal. This is contrary to common sense. There is nothing in the nature of the case which will explain the large differences in the numbers occurring at the units of length indicated. A study of the tables shows that the frequencies are concentrated on the even and the one half inches. No such concentration, however, actually occurs. The reason for the concentration is the wish of the one who did the measuring. Arbitrary units of length—a continuous fact—were set up, and then the numbers (a discrete fact) falling at *approximately* these lengths were identified.

The frequencies in column (a), although they appear to be precise and accurate, are in fact inaccurate. Neither in the world at large nor in the sample selected for measurement does such a condition as there indicated obtain. Indeed, greater accuracy from group to group and over the entire range of measurements is secured by expressing the frequencies in wider groups. This is done in columns (b), (c), (d), and (e). It is more correct to say, for instance, that 1152 cases were encountered measuring 10 to 11 inches in length than to say that 514 were 10; 61, 10¼; 532, 10½; and 45, 10¾ inches. The thing which distinguishes this distribution from that of the mortgage interest rates is the unreal concentration upon even and half inch units. In the former case, concentration actually exists and should be preserved; in the latter case, it is fictitious and should be smoothed out by widening the groups. This process in the former case sacrifices accuracy; in the latter, it helps to realize it.

In fixing the number, the widths, and the origin and termination of groups representing continuous series, the aim should be to (1) leave no group unrepresented by frequencies, (2) provide for a gradual distribution of the instances through the groups, (3) permit the frequencies gradually to reach a maximum and "tail off" to a minimum, and (4) have the widths exceed the differences observed in the measurements.

In frequency distributions, both of discrete and of continuous series, it is desirable to make the groups of equal width. If this rule cannot be followed because the use of equal sized groups (1) is too detailed for some and not detailed enough for other frequencies, (2) results in securing a distribution not properly descriptive of the frequencies over their entire range, (3) would leave some groups vacant, etc., then the larger groups should be multiples of the smaller ones. While the larger ones cannot be broken up, the smaller ones can be combined when comparisons are desired.

Table 19, showing the distribution of wage-rates of operators in woolen and worsted mills in the United States, illustrates the use of unequal groups and suggests the errors into which one may be led through their use.

By ignoring the widths of the groups and assuming them as equidistant—a likely thing to do unless one is accustomed to studying such data—it appears that the regular descending order of the frequencies for both males and the total is abruptly broken at the frequency 2604 for the total, and at 2109 for the males, thus giving a new point of concentration of the wage earners. The larger numbers of frequencies, of course, are due to the use at this point of wider groups. This table can only rightly be interpreted if full account is taken of the fact that the distribution applies to groups with limits of 2, 5, 6, 10, and 15 cents, as well as to one group which is open at the upper side. If the table had been properly constructed, the order of the units—hourly rates of wages—would have been inverted, and uniform size groups, or groups which are reducible to multiples of each other, used. When different

sized groups are used, breaks should be made in the body of the table to call attention to the fact.

In writing the limits of groups, a smaller fraction of the whole unit should not be used than was employed in the actual process of measurement. For instance, wages measured in cents should not be expressed in groups reading in fractional parts of a cent. Likewise, if measurements are made to the nearest half inch, the limits of the groups should not be indicated by quarter inches. Moreover, it is desirable, in order to guard against confusing the upper limits of a lower group with the lower limits of an upper group, to avoid writing the two in the same form. For instance, the group "30 to 40" should be written "30 but less than 40." In this form, it is clear that a frequency of 40 belongs in the group 40 but less than 50.

TABLE 19

FREQUENCY TABLE SHOWING THE NUMBER OF THE OPERATIVES IN WOOLEN AND WORSTED MILLS IN THE UNITED STATES, BY SEX AND BY HOURLY RATES OF WAGES

(*Report of the Tariff Board on Schedule K*, Vol. IV, part 5. House Document No. 342, 62d Congress, 2d session, p. 997)

| HOURLY RATES OF WAGES | TOTAL | MALES | FEMALES |
|---|---|---|---|
| Total | 30,454 | 17,343 | 13,111 |
| 75 cents and over | 33 | 33 | — |
| 60 to 74.99 cents | 60 | 59 | 1 |
| 45 to 59.99 cents | 109 | 106 | 3 |
| 35 to 44.99 cents | 291 | 287 | 4 |
| 30 to 34.99 cents | 486 | 451 | 17 |
| 25 to 29.99 cents | 2,004 | 1,849 | 155 |
| 20 to 24.99 cents | 2,604 | 2,109 | 495 |
| 18 to 19.99 cents | 1,682 | 1,142 | 540 |
| 16 to 17.99 cents | 2,635 | 2,036 | 599 |
| 14 to 15.99 cents | 4,926 | 3,729 | 1,197 |
| 12 to 13.99 cents | 6,007 | 3,186 | 2,821 |
| 10 to 11.99 cents | 6,153 | 1,453 | 4,700 |
| 8 to 9.99 cents | 2,722 | 757 | 1,965 |
| 6 to 7.99 cents | 661 | 133 | 528 |
| Less than 6 cents | 99 | 13 | 86 |

Table 20 illustrates a flagrant violation of these principles. The upper boundaries of the second and ninth groups are indefinite. According to the way in which they are stated, items of 3 and 21 per cent, respectively, are not to be included, yet it is certain from the succeeding groups that they are in-

TABLE 20

TABLE SHOWING THE PERCENTAGE RELATION OF THE ASSESSMENT OF PERSONAL PROPERTY TO TOTAL ASSESSMENT

(*Report of the Joint Legislative Committee of the State of New York*, Albany, 1916, p. 260)

| RELATION OF PERSONAL PROPERTY ASSESSMENT TO TOTAL ASSESSMENT | NUMBER | WIDTH OF GROUPS IN PER CENTS |
|---|---|---|
| Total ........................ | 53 | |
| Less than one per cent.............. | 2 | Less than one |
| From one to three per cent.......... | 5 | 3 * |
| From four to six per cent............ | 5 | 2 † |
| From six to eight per cent........... | 10 | 2 † |
| From eight to eleven per cent........ | 7 | 3 † |
| From eleven to thirteen per cent...... | 12 | 2 † |
| From thirteen to eighteen per cent.... | 5 | 5 † |
| From eighteen to twenty per cent..... | 3 | 2 † |
| From twenty to twenty-one per cent.. | 3 | 2 * |
| Greater than twenty-one per cent..... | 1 | Indeterminate |

* Upper limit included.
† Upper limit not included.

cluded. If they are, the order is an exception to that which characterizes the majority of the groups. As a result, one is left in doubt as to what is intended. Moreover, the groups are so different in size that discredit is thrown upon the whole table.

## XII. CONCLUSION

A detailed summary of this chapter seems unnecessary. The aim has been to consider only the most important aspects of the subject. The more general phases of classification and their bearing upon scientific method have for the most part

been taken for granted.[1] They need no extended consideration in this connection. We have striven only to show the application of classification to statistical facts.

The technique of tabulation has been approached with the problem of the statistician in view, the aim being to call attention to and to warn against certain indefensible practices commonly followed and at the same time to formulate, as nearly as can be done, rules of general application. Attention is drawn to the characteristic differences in statistical data and to the appropriate methods of showing them in tables. A logical background for the existence of tables, and the reciprocal relation of the point of view from which data are considered and the way in which they are presented in tables have been emphasized. Tabulation is always more than a mechanical drawing of lines and inserting of numerical symbols. To its purpose and technique, too much attention cannot be given.

### References

Bowley, A. L., *Elements of Statistics,* 4th Edition, King, London, 1920, Chapter IV, pp. 52-81.

Bowley, A. L., *An Elementary Manual of Statistics,* MacDonald and Evans, London, 1915, Chapter VI, pp. 50-56.

Day, E. E., "Classification of Statistical Series" in *Quarterly Publications American Statistical Association,* December, 1919, pp. 533-535.

Durand, E. D., "Tabulation by Mechanical Means," etc., in *The Transactions of the International Congress on Hygiene and Demography, 1912,* Section Nine, pp. 83-91.

King, W. I., *Elements of Statistical Method,* Macmillan, New York, 1912, Chapter IX, pp. 83-90.

Watkins, G. P., "Theory of Statistical Tabulation," in *Quarterly Publications American Statistical Association,* December, 1915, pp. 742-757.

Žižek, Franz, *Statistical Averages,* Holt, New York, 1913 (Translated by W. M. Persons), Chapter I, pp. 7-24; Chapter V, pp. 80-91.

[1] These are admirably treated in Venn, John, *Empirical Logic,* Macmillan, New York. 1907. and in *The Logic of Chance.* 3rd Edition Macmillan and Co., London, 1888; as well as in Jevons, W. S., *The Principles of Science,* 2nd Edition Macmillan and Co., London, 1920.

# CHAPTER VII

## DIAGRAMMATIC PRESENTATION

### I. Introduction

Amounts and frequencies are tabulated in Arabic or Roman numerals; they are illustrated by lines, bars, surfaces, volumes, and maps. The facts themselves may be either discrete or continuous, and be related to different times, different places, or to different conditions at the same time or place. The various devices used to illustrate discrete data are treated in this chapter under the heading *Diagrammatic Presentation.* Those used to illustrate continuous series are discussed in the following chapter, *Graphic Representation.*

In the chapter on *Classification—Tabular Presentation* the function of a logical classification of statistical data and of their arrangement in tables was discussed at length. It was learned that primary data must be classified and reduced to order from the heterogeneous form in which they are reported, while secondary data must be rearranged, separated, combined, and worked over to suit the purposes for which they are intended. Respecting both, the first essential to tabulation is classification. The classes into which data fall are arranged logically in the order of their importance, the data themselves being placed in the lines and columns of tables. Such an arrangement facilitates study, throws related things together, and suggests analysis. Our purpose in this chapter is to contrast tabulation with diagrammatic presentation, and to discuss the value of the various forms of illustration currently used for this purpose.

The purpose of tabulation is to reduce masses of facts to

logical order according to the units of measurement in which they are expressed and for the purposes desired. The functions of diagrams are to illustrate these facts according to the order worked out by tabulation. Tabulation is a condition of analysis; diagrams are generally illustrations of conclusions from analysis. The former is necessary in interpretation; the latter are useful in explanation and exposition. Classification and tabulation precede; the use of diagrams follows. The former clarify the meaning of data; the latter frequently obscures it. Diagrams can never displace tabulation; they may conveniently accompany it if used with discretion. Tabulation alone suggests study and analysis; diagrams alone are more likely to serve as bases for conclusions arrived at without study, and to foster a disregard for the details from which diagrams are drawn. Careful analysis of tabulated data is frequently necessary before their full meaning is divulged; a superficial view of diagrams is often gathered from mere inspection.

Diagrams rarely add *new* meaning to facts which they illustrate. What they do do is to *add to* the meaning by throwing it into relief and by clarifying it.

It is unwise, as a general rule, to use analogies, but one may be hazarded in order to show the dependence and secondary character of diagrams in statistical studies. Botanists, in classifying plants, use established points of distinction to separate them into groups. The common characteristics are noted in detail and become the bases for further classification, each sample or group of samples being differentiated from the others by the presence or the absence of chosen criteria. Groups and sub-groups are distinguished and these again are studied in the light of the distinguishing marks chosen. This process is continued until the points of difference are exhausted, or until some scheme of organization extending throughout the whole group or groups is discovered. The methods of classifying plants are analogous to those of classifying statistical data. The common characteristics become the cri-

teria of distinction. Labeling, naming, and mounting botanical specimens are processes analogous to illustrating and "mounting," by statistical diagrams, the relations established through tabulation. The former may exist and be independent of the latter in both instances; the latter grow out of and are conditioned by the former in all cases.

What has been said is not meant to detract from the value of diagrams as *aids* in statistical studies. Its purpose has been solely to show that they are subordinate to classification and tabulation. Diagrammatic illustrations of data can never replace the data themselves, no matter how accurately they tell the truth nor how skillfully they are drawn. They are at best statistical *aids* and should be so considered by those who use them. A well-drawn and cleverly constructed diagram is never a guaranty of the value of the statistical facts which it illustrates.

This contention is supported by a review of the *Statistical Atlas of the United States*. The reviewer, in questioning the need of such a volume, raises the point whether it is desirable to segregate the illustrations from the tables and text analysis. He says:

"Is the policy of segregation a wise one? Presumably these maps and diagrams have had and will continue to have their most effective use in connection with the tables and text with which they were originally published. To place them in a separate volume with the barest textual comment seems unduly to burden the graphic method of presenting facts. Frequently charts and maps greatly strengthen the textual exposition of a subject; they seldom serve as a complete substitute for editorial analysis."[1]

The psychology of the use of statistical diagrams is worthy of brief consideration. It is difficult to hold in mind a great mass of figures. Relations are likely to be obscured in the effort to remember the amounts themselves. Well-constructed tables, however, partly compensate for this limitation. But

[1] Day, E. E., Review of "Statistical Atlas of the United States," in *The American Economic Review*, September, 1915, pp. 648-650, at p. 650

even when facts are arranged in tabular form, the size of the items, in all but summary tables, is given chief emphasis. But size is seen in its absolute rather than in its relative aspects. Degrees of difference between items at the same time, the same place, and for different times and places are not easily comprehended when data are expressed in quantities. The order in which they are arranged may in part compensate for the limitations of tabulation, but it cannot entirely overcome them. If, for instance, an order of arrangement is according to magnitude or frequency, as when districts are arranged in the order of amount or number of sales; or if it is consecutive, as when loans are listed according to size of interest rates, an idea of extreme change is readily grasped. The distribution, amount, and frequency of change, however, are emphasized when they are thrown into relief by some form of diagrammatic illustration. On the other hand, when no definite order in tabulation is followed, or when the order of arrangement is illogical—or, if logical, is not consistently followed—differences in time, space, and frequency do not stand out.[1] It is to overcome these imperfections and limitations of tabular arrangement, to introduce devices for showing the proportional relations between facts, and to emphasize the relations of amounts to space, that diagrams of various types are used.

The power of visualization is only partly realized in tabulation. True, if tabular forms are properly drawn, data are arranged in lines and columns according to a logical plan. But relations do not stand out. They may be worked out by means of percentages and ratios, but such expressions are difficult to visualize. Absolute and relative differences in interest rates on real estate mortgage loans in Illinois, for instance, may be compared with the frequencies with which the various rates occur, but it is not easy to relate the rates geographically to the counties of the state without using a

[1] The desirability of having every tabular form determined according to a definite plan and follow a logical order is developed in the preceding chapter, pp. 132-136.

statistical map. A tabular form in which the counties are arranged alphabetically may have no logical significance. To group the counties by rates may not necessarily be to include contiguous territory. Where space relations are involved, statistical diagrams help to make them clear. Even where geographical distribution is not important, they help to show relations, proportions, and sequences.

Probably sufficient has been said to indicate in a general way that diagrammatic illustration adds something to tabulation. Just how this is done and in what way by different types of diagrams will be made clearer by a discussion of the different forms used, the technique of their construction, and the psychological basis upon which each rests.

## II. Diagrams for Illustrating Frequency or Magnitude Alone

### 1. alternative types—good and bad features of each

The diagrammatic forms commonly used to illustrate amounts and frequencies which are discrete are lines, bars, surfaces, and volumes. As a class, these are called *pictograms.*

Suppose certain data were available concerning the stocks of merchandise of a retail store. The amounts on hand at dates of inventory for a succession of years constitute a dis-

TABLE 21

Stocks of Merchandise Illustrating Different Types of Statistical Series

| (Time Series) | | (Space Series) | | (Condition Series) | |
|---|---|---|---|---|---|
| Years | Amounts on Hand at Date of Inventory, Jan. 31 | Departments | Amounts on Hand at Date of Inventory, Jan. 31, 1924 | Methods of Taking Inventory | Amounts on Hand at Date of Inventory, Jan. 31, 1924 |
| Average | $210,000 | Total | $180,000 | Total | $180,000 |
| 1921 | 200,000 | A | 60,000 | At Cost .......... | 30,000 |
| 1922 | 240,000 | B | 40,000 | At "Market" ...... | 110,000 |
| 1923 | 220,000 | C | 30,000 | At Appreciated Value | 5,000 |
| 1924 | 180,000 | D | 50,000 | At Depreciated Value | 35,000 |

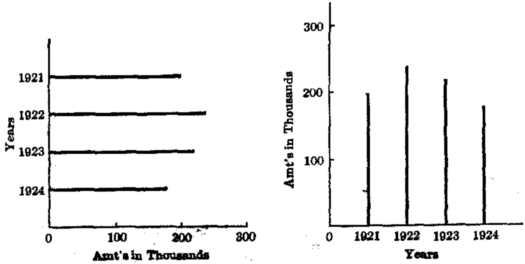
crete time series; the amounts of stock classified by departments, a discrete place series; and amounts classified by the methods of taking the inventories, a discrete condition series. The data are in Table 21.

To illustrate each of these series, various forms of diagrams may be used, the parts standing for and being proportional to the amounts. If lines are used for the time series, for instance, the amounts may be shown as in Figure 4, a

FIGURE 4

horizontal arrangement being used and the lines having no common base. On the other hand, the points of origin may be made the same in all cases, the lines extending either horizontally or vertically. The diagrams, respectively, would then appear as in Figure 5. In place of lines, *bars* of equal width—broad lines, in fact—may be drawn vertically or hori-

FIGURE 5

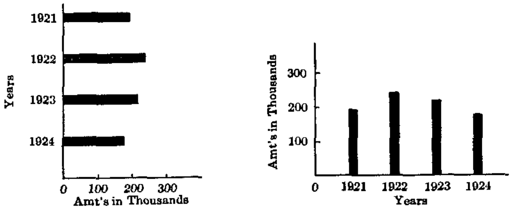


zontally with or without a common base. Horizontally drawn without a common base they would appear as in Figure 6;

*FIGURE 6*

horizontally and vertically with a common base, in these forms, respectively. An alternative method of illustrating the

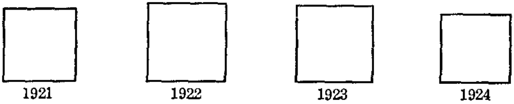
FIGURE 7




same series is to use surfaces in some such fashion as in Figure 8; or as in Figure 9. Cubes also may be employed. The

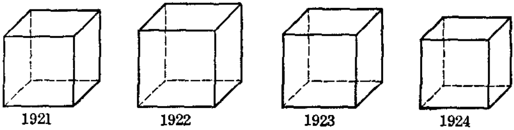
*FIGURE 8*

FIGURE 9




illustrations, if horizontally placed, would appear somewhat as follows:

FIGURE 10

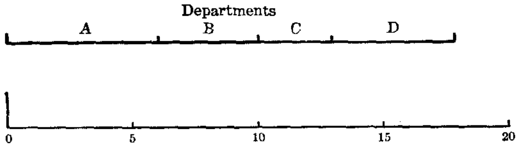



The facts shown in Table 21 are discrete and separate. Neither the times, the places, nor the conditions are dependent upon each other. While the amounts by years, by departments and by *methods of taking inventory constitute series, they are* unrelated to each other. They are separate identities. Moreover, because of the fact that *relative* size alone is illustrated, the lines, bars, surfaces, and volumes may have any dimensions desired, the only condition necessary to their faithfully illustrating the facts in question being that proportionally they bear the same relation to each other.

The same types of diagrams may also be used to illustrate the component parts of a total. For instance, if it were desired to make a diagram of the components of the total inventory on hand January 31, 1924—distinction being made by

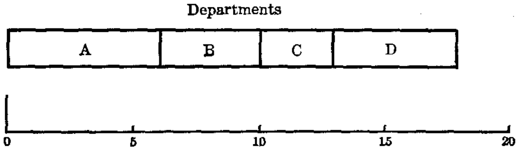
departments—lines, bars, surfaces, or volumes could be used. The line type would appear broken as in Figure 11. If the

FIGURE 11




bar type were used, it would appear in the form shown in Figure 12. The length of the bar is equal to the total inven-

FIGURE 12

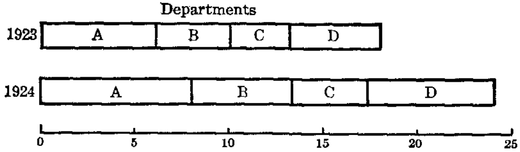



tory, and the lengths of the parts, to the amounts found in the different departments. The portion in Department A may be directly compared with the total because both have common points of origin. Those in Departments B, C, and D cannot be easily compared with each other or with the total because they do not have a common base. In this respect they are similar to the lines and bars, placed horizontally, which illustrate the inventories on hand in the different years.

If bars are used to show component parts at two different times or places, or under two conditions, then they will appear

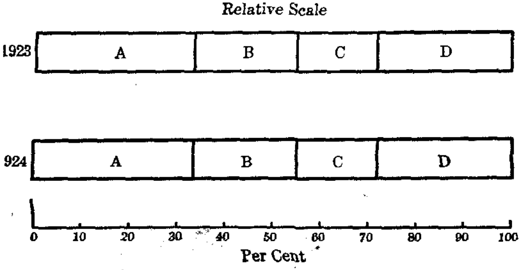
*in the following form—Figure 13—*different years being used for illustration. In this case, the respective lengths of the bars and of their component parts illustrate actual amounts. Components "A" and the totals in both years may be directly

FIGURE 13




compared with each other because they have a common base, and one dimension—horizontal—only is used. If the same facts were shown on a *relative* scale, the diagram would appear in the form shown in Figure 14. That is, the *total* inventory values at the two periods, while quantitatively different, are

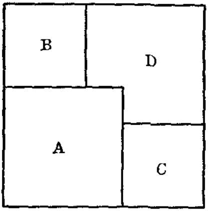
FIGURE 14



treated as equal in distributing on a proportional basis the amounts in the different departments.

Areas are sometimes used to show component parts, but their use is *not recommended.* Suppose it were desired to show by departments the components of the inventories and areas were used. The figure would appear somewhat as Figure 15, the total area equalling the complete inventory and the small areas the respective parts. None of the sections are directly comparable with each other or with the total—there is no common base. Moreover, since areas are used, the quantities are equal to the products of the sides of their respective rectangles, and cannot be readily compared.

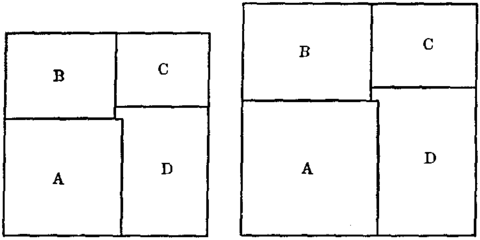
FIGURE 15




If surfaces or areas are used to show component parts at two different times or places, or under two conditions, then, using the illustration shown by bars, the figures would appear as in Figure 16. In such figures, the dimensions of the total areas, as well as of those of the parts, vary as the square roots of the surfaces. Comparisons in such cases are extremely difficult if not impossible. *Figures of this type should not be used.*

Circles or pie diagrams are also used to show component relations. *For this purpose, they are not recommended.* If the component parts of the total inventories at a given time, dis-

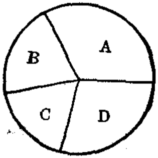
FIGURE 16




tributed according to departments, are illustrated in this way, the resulting figures would be as shown in Figure 17.

The total area represents the total inventory: the areas of the parts, the amounts in the respective departments. *Areas* are used in all cases. But the area of a circle is secured by squaring the radius and multiplying by $\pi$ — 3.1416. When it is divided into components, the parts *appear* to stand in the relation of their respective chords. But this is not the case, since the smaller the sector, the longer the chord relative to its corresponding arc, and vice versa. The areas of the sectors are proportional to their respective arcs, but not to their respective chords. But it is the arcs which cannot be easily

FIGURE 17

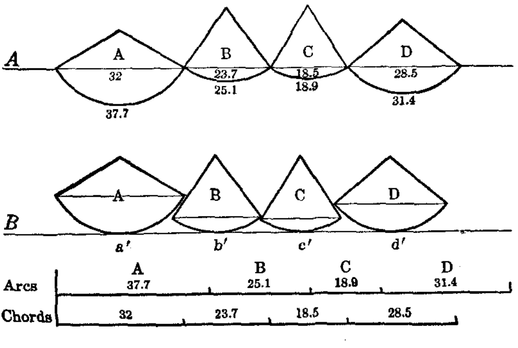


compared—they are circular—and relative lengths are not apparent. To be compared, they must be straightened out in the mind. The ease with which this can be done varies inversely with their length.

All radii of a circle, of course, are equal and the lengths of the arcs are proportional to the angles at the center. But it is as difficult to compare the relative sizes of the angles as it is the lengths of the arcs.

The types of figures which one is asked to compare, when sectors of circles are used to show component parts, are illustrated in Figure 18. In the part marked "A" the chords are placed in a straight line. It is apparent from the illustration that the areas of the sectors have little relation to the chord lengths, and yet it is these which attract the eye in the pie diagram. In the part marked "B," tangents, in the form of a continuous straight line, are drawn to the respective sectors at points $a'$, $b'$, $c'$, $d'$, the sectors having been separated. The areas of these figures cannot be readily compared—they are not graphic. The lower part of Figure 18 shows the respective

FIGURE 18

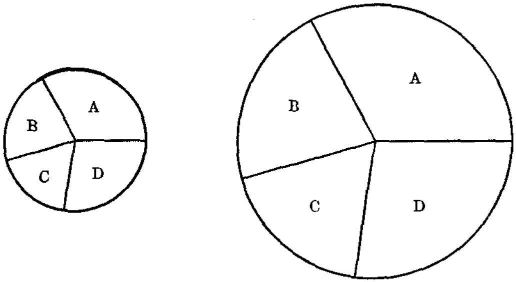


lengths of the arcs and of the chords with the differences between them. The larger the angle, the greater the difference between the chord and the arc, and vice versa.

A pie diagram is a clumsy and defective method of illustrating component parts; a bar of uniform width—that is, a one-dimensional figure—is much more satisfactory.

The use of circles or pie diagrams to show component parts of things at different times, different places or under different conditions, is even less defensible. Such an illustration as Figure 19 is sometimes used for this purpose.

FIGURE 19



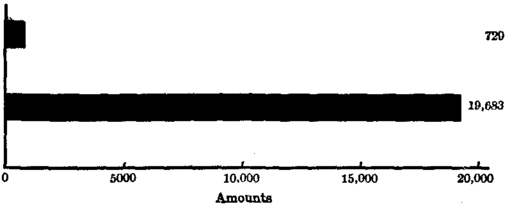

It is necessary, in case actual amounts are used, to compare (1) the sizes of two circles, (2) the proportions of each taken up by the different parts, and (3) the comparative sizes of the parts in one with the corresponding parts in the other. This is asking too much; it cannot be done. For the eye to compare the areas of the different parts in the *same* circle is difficult enough; but to compare the relative areas of corresponding parts in two circles whose total areas vary as the squares of their radii is impossible. Concerning the disadvantages of the pie chart, a recent writer says:

"It is worthless for study and research purposes. In the first place, the human eye cannot easily compare as to length the various arcs about the circle, lying as they do in different directions. In the second place, the human eye is not naturally skilled at comparing angles—those angles at the center of the circle, formed by the various rays or radii and subtending the various arcs. In the third place, the human eye is not an expert judge of comparative sizes of areas, especially those as irregular as the segments of parts of the circle. There is no way by which the parts of this round unit can be compared so accurately and quickly as the parts of a straight line or bar."[1]

Amounts, frequencies, and component parts cannot be readily illustrated by cubical figures, the contents of which vary as the cubes of their dimensions. Two quantities such as 729 and 19,683, for instance, are illustrated by the use of bars—one dimension being used—in Figure 20. Cubes showing the same facts are given in Figure 21. That is, the respective dimensions stand in the relation of 9 to 27, or 1 to 3, and the contents as 729 to 19,683, or 1 to 27. It is not easy to think in terms of three dimensions; by the casual reader, volumes are read in one dimension.

Component parts are even more difficult to show by the use of volumes. In order to determine the dimensions to be used,

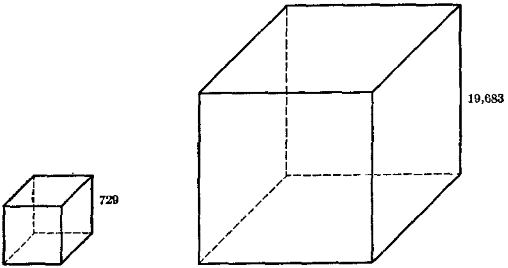
FIGURE 20




[1] Karsten, Karl G., *Charts and Graphs*, Prentice-Hall, New York, 1923, p. 91.

it is necessary to take the proportionate parts of the respective contents, and to extract their cube roots. The resulting figures are very confusing; they are *not* graphic.

FIGURE 21




## 2. EXAMPLES OF STATISTICAL DIAGRAMS IN CURRENT USE

Various types of diagrams illustrating discrete series are given in the following pages. Because of the lack of space and the fact that the discussion does not purport to be a treatise on diagrammatic presentation, only a few kinds are introduced. The interested reader may consult with profit the books which deal more fully with this topic, reference to which will be found at the close of this chapter.

Figure 22 shows the relative prices of a number of farm products by years, the articles being distinct and the prices unrelated to each other. Separate bars properly illustrate the respective relative prices. To have connected them by lines would have given an incorrect impression; it would have made it appear that the relative heights were in some way dependent upon each other. The diagram, moreover, shows a break in

the time units in which the prices are shown. Data for the years 1915 to 1919, inclusive, are missing. Accordingly, attention is called to this fact by the white area between 1914 and 1920. Figure 22 illustrates a discrete series in time.

FIGURE 22

DIAGRAM SHOWING DISCRETE TIME SERIES

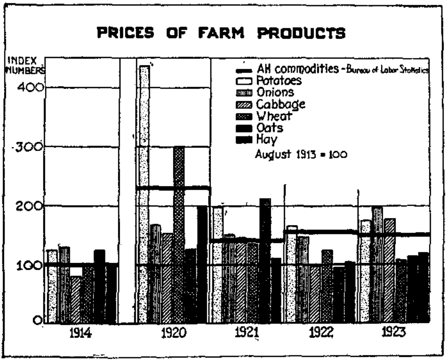



Figure 23 shows a discrete space series, horizontal bars being used to show per cent changes in 1923 over 1921. The order of arrangement is descending. Inasmuch as the facts are discrete, the bars are distinct and evenly spaced. The "grand total" (in fact an average) is removed from the detail by a slightly wider space than that used to separate its parts.

Figure 24 shows another discrete space series. In this diagram, the areas having an excess of exports are listed in descending order, and those having an excess of imports in ascending order. The total appears at the bottom of the diagram, removed from the details.

FIGURE 23

DIAGRAM SHOWING A DISCRETE SPACE SERIES

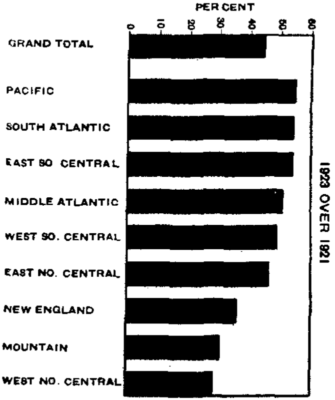



FIGURE 24

DIAGRAM SHOWING A DISCRETE SPACE SERIES

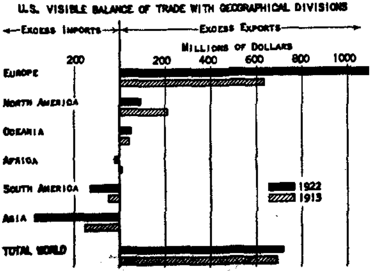
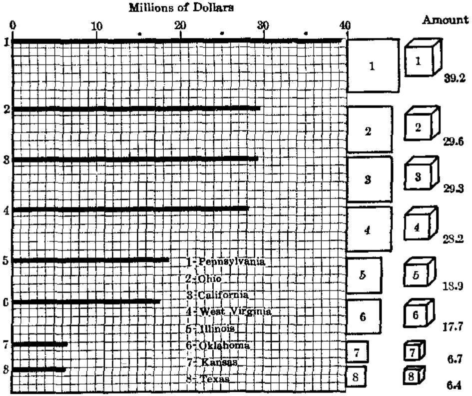


Figure 25 shows how a discrete space series may be illustrated by bars, surfaces, and volumes. Absolute and relative differences are much more apparent in the bars than in either of the other forms of illustration. Both may be verified by inspection when one dimension is used; when two and three dimensions are employed, however, they can be verified only by computation. The surfaces vary as the squares, and the volumes as the cubes of their dimensions.

FIGURE 25

VALUE OF PETROLEUM AND NATURAL GAS, BY STATES, 1909

(Illustrations of Lines, Surfaces, and Volumes)



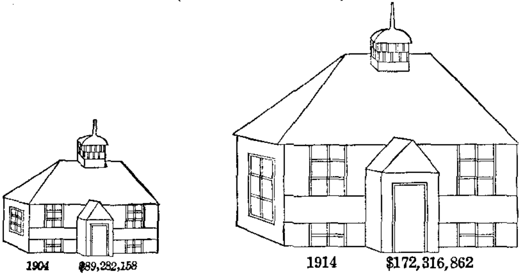

Figures 26 and 27 show solids drawn out of proportion, thus giving erroneous impressions. Such figures are meant to be helpful, but they are confusing and absurd. In Figure 26, absolute amounts for 1904 and 1914, respectively, stand in the relation of 51.8 to 100. The illustrations show them to be 12.5 to 100. In Figure 27, the relation between the amounts

is 44.3 to 100; the diagrams show it to be as 6.42 to 100. In both cases, fortunately, the amounts accompany the diagrams, and the errors can be corrected.

FIGURE 26

PUBLIC SCHOOL PROPERTY IN 1904 AND 1914

(Solids drawn out of Scale)

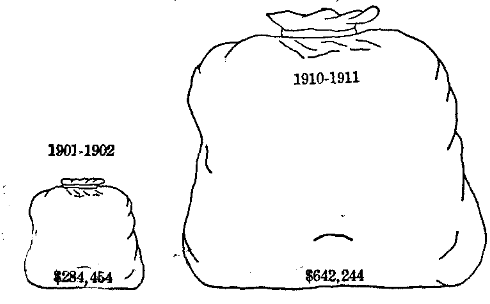



Per Cent of Increase in 10 Years, 93%

FIGURE 27

PAYMENTS, ACCOUNT BONDED DEBT AND INTEREST, ON COUNTY BONDS

(Solids drawn out of Scale)



A discrete condition series is shown in Figure 28, a descending order (except for the miscellaneous item) in cost per employee being used. The different industries are separated by equal spaces, the bars being distinct. The average is placed at the bottom of the illustration, is removed from the detail, and indicated by a distinct type of shading. The diagram ought to have a scale and contain the amounts in tabulated form.

The bar showing the cost per employee in mining is left jagged at the end, thus calling attention to the fact that the precise amount is not shown.

FIGURE 28

DIAGRAM SHOWING A DISCRETE CONDITION SERIES

*(Industrial Medical Departments:*

Average Annual Cost per Employee, by Industry)

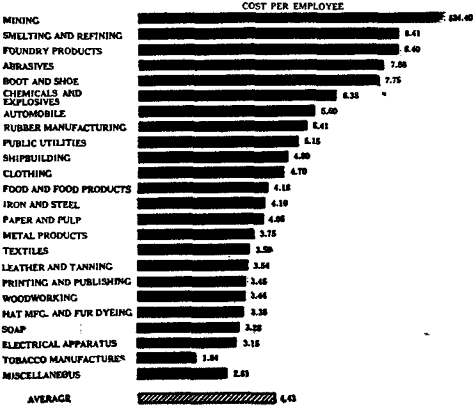


FIGURE 29

Diagram Showing Component Parts—Discrete Time Series

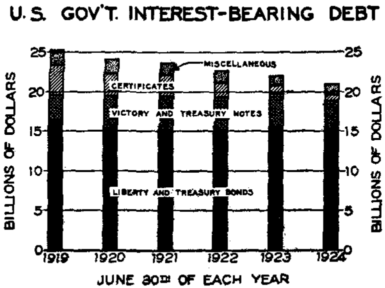
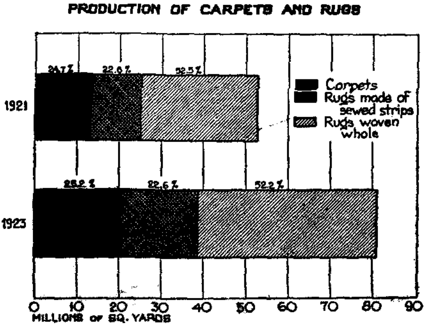



FIGURE 30

Diagram Showing Component Parts—Discrete Time Series



Bars are used in Figures 29 and 30 to show component parts in discrete time series. In Figure 29, the arrangement of the bars is vertical, the parts being expressed in quantities; in Figure 30 the arrangement is horizontal, both amounts and proportions being given. In both cases, since the facts are discrete, the bars are distinct and separate.

The uses to which circles or pie diagrams are put in illustrating component parts of a whole at a given time, relative proportions at different times, and different amounts and proportions at different times were discussed above. The following diagrams are illustrative of those being used.

FIGURE 31

PIE DIAGRAM SHOWING COMPONENT PARTS

# The Edison Dollar of Income

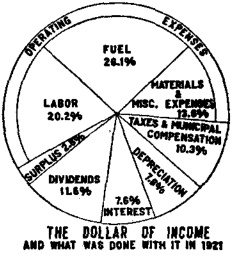



Figure 31 shows the distribution of a dollar of income received in 1922 by the Commonwealth Edison Company, Chicago, the total area of the circle being 100 per cent, and the different segments proportions of the total.

On the other hand, Figure 32 shows the proportions which the important items of a family budget constitute at different times. For purpose of distribution, the total budgets are shown to be equal, the areas of the circles being the same. The segments are proportionally but not quantitatively comparable.

FIGURE 32

PIE DIAGRAMS SHOWING COMPONENT PARTS

*(Percentages of Expenditures for Major Items of Family Budget)*

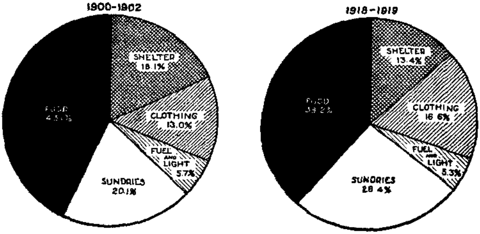
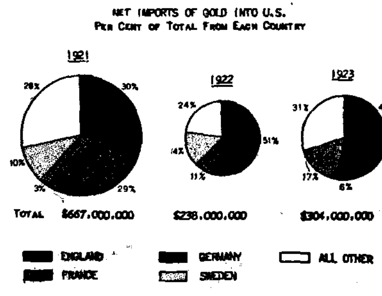



FIGURE 33

PIE DIAGRAMS SHOWING COMPONENT PARTS BY YEARS

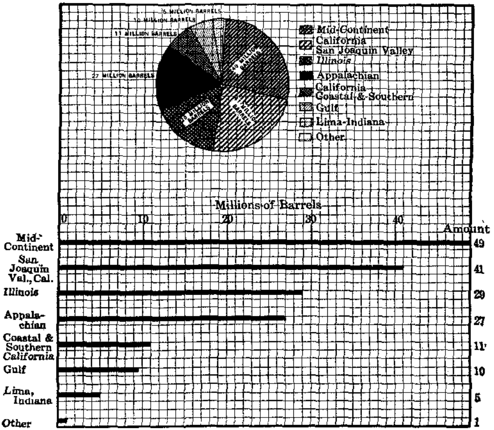


In *Figure* 33, *amounts* varying from year to year are shown by the *areas* of circles. Each separate amount is then divided into its component parts, these being indicated as proportions of the total. It is difficult to interpret such diagrams. For instance, the white *area* — "all other" — in 1923 is smaller than the corresponding area in 1921, although proportionally it is

FIGURE 34

PRODUCTION OF PETROLEUM, BY FIELDS, 1909

(Sectors of Circles and Lines)



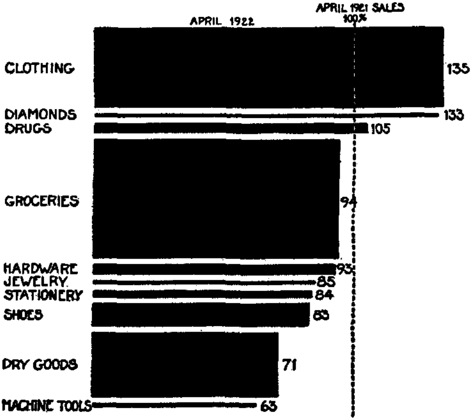

larger. Similar observations apply to other segments. The different parts of the total area in *any* year are directly comparable; the same parts in different years are not directly comparable. Bars either *vertically* or *horizontally* placed bring out the relations much better than do circles.

The use of bars and circles to illustrate the same facts are contrasted in Figure 34.

In some cases, bar diagrams, varying in two dimensions, are used to illustrate discrete facts. This is done in Figure 35, which shows horizontally, by the length of different bars, the relation of sales in April, 1922, to sales in April, 1921; and vertically, by the widths of the bars, the relative amounts presumably sold in April, 1922.[1]

FIGURE 35

TWO-DIMENSIONAL BAR DIAGRAM SHOWING DISCRETE CONDITION SERIES



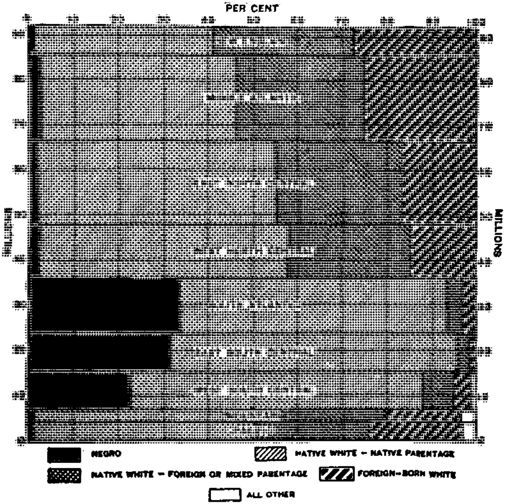

A similar bar chart using two dimensions is illustrated in Figure 36. The interesting thing about this figure is that absolute amounts are shown by widths of bars, lengths in all instances being identical and constituting 100 per cent. By *cross-hatched surfaces not only are geographical divisions, but*

[1] So far as the form of the chart is concerned, the relative amounts could be those in either period.

color, race, nativity, and parentage shown for the population of the United States. The figure admits of being read in two dimensions the same as a table, yet no confusion results.

FIGURE 36

Color or Race, Nativity, and Parentage, by Divisions of the United States, 1910

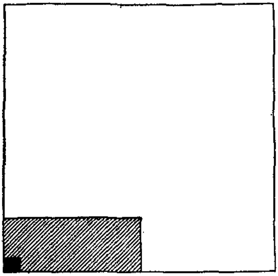



Occasionally, but fortunately not often, surfaces within surfaces are used to show a total and its component parts. An example of this atrocious practice is shown in Figure 37. In commenting upon this diagram, the writer using it says: "The large area represents the approximate annual business of wholesale druggists of the United States—in round numbers

$400,000,000. The shaded area represents the proportion of credits to total sales. The black area in the lower left-hand corner shows losses on total credits." To this illuminating (?) statement the reader will instinctively ask: What is the proportion of credits to total sales? What proportion of total credits are losses? These questions cannot be answered because (1) no data accompany the diagram, and (2) no one will take the trouble to *compute* the proportions from the diagram. Such illustrations are worse than useless.

FIGURE 37

Two-Dimensional Diagram Showing Components by Use of Surfaces Within Surfaces



The foregoing diagrams, as said above, are illustrative of certain types in current use.

### 3. GENERAL RULES TO BE OBSERVED IN THE USE OF STATISTICAL DIAGRAMS

The need for following a logical and consistent order of arrangement is equally as important in illustrating statistical facts as in tabulating them. For instance, when dealing with geographical distributions, where contiguity of districts is important, this order should be followed. Where time is a factor, it should control the arrangement. As a rule, less at-

tention is paid to the order of arrangement in illustrations than in tabulations because violations are not so apparent. False impressions are given by using an illogical order and by omitting all concrete data. Deception, if willed, is not difficult to effect. The apparent is easily confused with the real. It must be remembered that in the case of diagrams it is the eye and not necessarily the intellect to which appeal is made. In this fact lies the chief source of danger in the tendency to think exclusively in terms of illustrations.

Diagrams of whatever type should be accompanied by the data which they illustrate. When this is done, the two supplement and correct each other. The suggestive power of diagrams is not interfered with, and at the same time precaution is taken against the tendency to place reliance in them alone. Moreover, the failure to include concrete data may not then be used as a partial justification for drawing false conclusions. The data not only serve as a record of the thing illustrated but also as a test of the accuracy of the illustration.

When lines or bars are used, their widths generally have no significance. Sufficient space between them should be allowed so that they will appear distinct. It is necessary, however, when data are classified into unequal-sized frequency groups to use lines of different widths. In such cases, it is the surfaces and not the linear dimensions which are important. The widths of lines or bars will then vary with the widths of groups, but this will not be confusing provided the ordinate scales are properly indicated, and the surfaces are interpreted in terms both of length and breadth. To depend on abscissa scales alone is not sufficient. It is this error which often explains the misinterpretation of data so grouped. An illustration of the erroneous conclusions to which people may be led by failing to take into account the changing sizes of groups is given in a recent study of the national income tax returns.[1] This failure is common and the reader should be

[1] See Falkner, Roland P., "Income Tax Statistics," *Quarterly Publications of the American Statistical Association*, June, 1915, pp. 523, 537.

constantly on the lookout for it when he is interpreting statistical diagrams.[1]

Confusion frequently results from including too much in a single diagram, the complexity of detail in whole or in part defeating the purpose which it is intended to serve. It is well to keep in mind the general rule that for diagrams to be effective they must be simple and easily understood. Complex relations can generally be more adequately shown by tables than by diagrams. In some cases, however, even for those which are relatively complex, diagrams are helpful because a number of comparisons can be made simultaneously. For those who are not accustomed to making and interpreting diagrams, however, it is wise to be conservative respecting the amount of detail which is crowded into them. There is no general and infallible rule respecting this matter, however, since much depends upon the idea which one wants to emphasize, the type of diagram used, the size of illustrations, the skill with which they are drawn, the consumers to whom they are addressed, etc.

In summarizing the discussion of the use of diagrams in illustrating statistical facts, attention should be called to the appeal which such figures make to the eye, and to the ability which they have to make plain relations and sequences which in tabular form remain abstract. For instance, a hundred per cent becomes significant in a line of a definite length. Likewise, any proportion of this amount is vividly represented by a line somewhat shorter than the one which represents the whole. Undoubtedly, when both quantities and illustrations are used, there results something additional to that which comes from using either alone. It is this something which has its basis in the psychological truth that the intensity with which a thing is perceived varies directly with the number of channels through which it makes its appeal.

[1] See illustration in *Report No. 4, Industrial Commission of Ohio* on "Industrial Accidents in Ohio, January 1 to June 30, 1914," Columbus, Ohio, 1915, pp. 36-37.

## III. Diagrams for Illustrating Frequency or Magnitude in Relation to Spatial Distribution

### 1. THE PSYCHOLOGICAL BASES FOR THE USE OF STATISTICAL MAPS

In order to show the relations between magnitude or frequency and geographical distribution, various types of statistical maps are used. As a class, they are known as *cartograms*. It is of interest briefly to discuss the psychological bases upon which their use depends, and to examine the different types currently employed.

The chief function of statistical maps is to show graphically amount or frequency in relation to position or space. For this purpose they are more satisfactory than tables. Data may be spread out geographically and amounts and frequencies studied in their relative and absolute aspects. Maps, moreover, are better suited for this purpose than are pictograms. Comparisons can be made of magnitudes in relation to position. The places of absolute and relative concentration and dispersion, together with the amount and rapidity of change from district to district, near and remote, are thrown into relief. Similar comparisons are difficult, if not impossible, from tabulations alone. The order of arrangement in tabulation, even if logical and consistent, is fixed. Inspection and study may suggest a different order from the one chosen, but rearrangement is possible only by retabulation.

The order in which data are illustrated on maps, while determined by amount or frequency—varying shades of color or density of cross-hatching, etc., indicating varying frequencies —is actually that of contiguity. It is, however, not inelastic. Comparisons may be made between remote as well as between contiguous districts. Similarities and differences stand out. They are shown not only alone and in relation to other amounts, but also as to positions. It is the introduction of the spatial concept which gives maps an advantage over tabular forms and simple pictograms. A new fact is represented—the

fact of position. A contiguous order may be followed in tabulation, but it lacks the concreteness which the projection upon a map gives it. The use of statistical maps makes it possible to visualize positions.

Maps show magnitude and position in different ways, depending upon the manner in which they are drawn, and the nature of the data which they represent. The different types, with their respective merits and demerits, are discussed below.

While maps are superior in many ways to tabulations, after all, they are secondary and simply illustrative. Classification of data precedes their illustration on maps. Illustration is dependent upon the order, range, and magnitude revealed through tabulation. In this respect, they are not different from pictograms. They do not stand alone. They support and illustrate statistical facts but do not displace them. Hence, they should be accompanied by concrete data, and be interpreted in terms of the units of measurement in which they are expressed. Not infrequently, all that can be done is to show the groups into which amounts characteristic of districts fall. If they are wide and the amounts dissimilar, it is impossible even to approximate exact frequency. To guard against any misunderstanding of what is shown, it is essential that the data should accompany the map. Their presence makes less likely hasty generalizations from appearances, and tends to direct attention not only to the map which serves to give an impressionistic view, but also to the data themselves. In the absence of the facts, different schemes of illustration may suggest radically different superficial interpretations, since not all types of maps are equally well suited for all purposes. Choice is not a matter to be treated lightly; it is to be determined by the nature and distribution of the data, the size and character of the groups into which they fall, the number of facts to be illustrated, etc. Maps, like simple pictograms, are aids in statistical presentation, but they are not indispensable in statistical analysis.

## 2. TYPES OF STATISTICAL MAPS

Statistical maps are of three general types: (1) those in which frequency is illustrated by different colors or by different shades of the same color; (2) those in which different shades of cross-hatching are used, the frequency or magnitude being indicated by relative densities; and (3) those in which various types of dots indicate frequency.

### *(1) Colored Maps*

The cost of making colored maps makes prohibitive their general use. Moreover, when the groups into which data fall are numerous, it is often easier to show gradual changes by varying the shades of black and white than it is by using separate colors or different shades of the same color. The use of different colors accentuates abruptness of change from one condition or district to another. Where different shades of the same color are used, it is frequently difficult to distinguish between them unless numbers or letters or some other identification marks are used. If color combinations are used, they should be complementary, the shades changing in harmony with the facts represented. Lighter colors and shades should represent one extreme; darker colors and shades, the other extreme.

On the use of colored maps, the following observation is of interest.[1]

"It is a cardinal principle in graphic representation that the visual impression should correspond directly to the facts as related to one another. Any scheme of color, therefore, which is not entirely logical, in a visual sense, is worse than misleading when applied to phenomena which are to be represented in a graduated series. A map in which the green, red, yellow, and blue are indiscriminately used to represent different grades of intensity of suicide, for example, is fully as difficult to interpret as the statistical tables which it is intended to elucidate. The only opportunity for representation by

[1] Ripley, W. Z., "Notes on Map Making and Graphic Representation," *Quarterly Publications of the American Statistical Association*, Vol. 6, 1898-1899, pp. 313-327, at pp. 314-315.

means of *unrelated* colors is offered in the case of such phenomena, for example, as the distribution of different nationalities or religions within a country where no relationship in point of fact between the several elements exists. . . .

"If colors are to be used at all, they should either be confined *to different intensities of the same color, or else, if the number of* shades be too great, two colors, red and blue, for example, may be employed, the deepest tints of each standing at the extremes of the series, and each shading down to an almost white color where the two join at the median line."

Excellent examples of colored maps may be found in the *Statistical Atlas of the United States,* published by the United States Census Bureau. Those who have occasion to use or interpret such maps should study them in relation to the choice of shades and colors, the varieties of uses to which they are put, the readiness and facility with which they may be interpreted, etc.

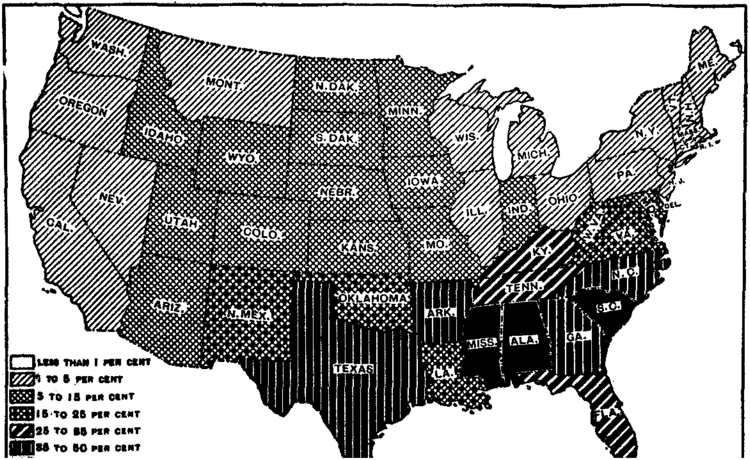
### (2) *Cross-hatched Maps*

The second type of maps is that in which some form of cross-hatching is used to indicate amount or frequency. Figure 38 is illustrative. Shades may range from white to black, extremes in the range of the thing represented being illustrated by extreme shades, and the condition which is more common, typical, or characteristic by medium shades. The number of shades to be used depends upon the number of groups into which data are divided. As in tabulation, groups should be of uniform size, shades representing equal ranges of units of measurement, rather than equal frequencies with which units occur. The number of times a shade is used in map making, as the frequency with which groups are encountered in tabulation, depends upon the total frequencies and the number of shades and size of the groups chosen. As widths of groups in frequency tables, so units of shades in maps should be uniform. When this rule is followed, choice of shades is of minor consideration.

FIGURE 38

PROPORTION OF MALES 10 TO 13 YEARS OF AGE ENGAGED IN GAINFUL OCCUPATIONS, BY STATES, 1910

(Cross-hatched Map)



The foregoing discussion applies primarily to the representation of a statistical series. Where unrelated and dissociated facts are illustrated, as, for instance, the number of consumers of a given commodity by districts, unrelated shades may be used. In such cases choice is determined largely by the desire clearly to contrast contiguous territories, and at the same time to bring out the detail necessary to the purpose in mind.

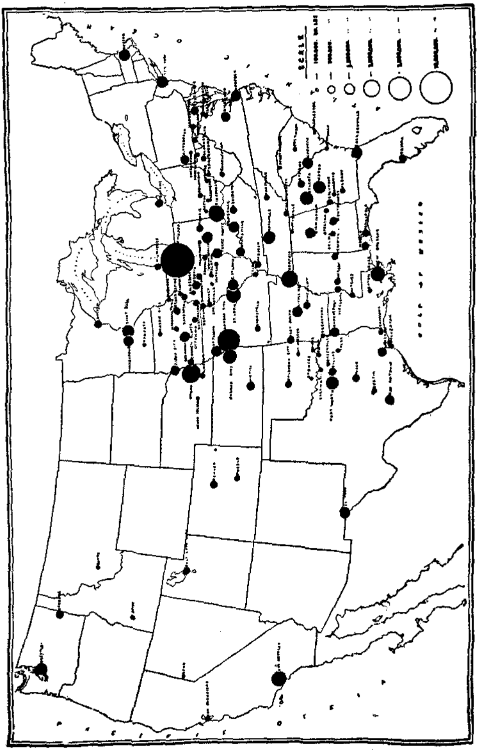
Both color and cross-hatching schemes are restricted to data of a "discrete" character. Where district boundaries mark complete changes, the presence or the absence, or the arbitrary limits to the operation of a thing illustrated, as do county or state lines for rates of increase of population, banking facilities, for instance, changes from district to district appear abrupt and violent. Such maps give the impression that absolute uniformity prevails within districts, and that changes occur only between them. For instance, maps illustrating, by districts, the per capita sales of merchandise; rates of changes in farm values or crop acreage; the average number of revenue passengers on street and electric railways per inhabitant, etc., must of necessity show uniform conditions within each district. Breaks appear only at boundaries. Division lines are predetermined. Such maps are "discrete" or broken. They should be used to illustrate only discrete series. When it is as necessary to show distribution by position within districts as it is between districts, that is, when the series illustrated are truly continuous, such maps give erroneous impressions. A more satisfactory method of illustration of both magnitude and frequency is then found in the so-called "dot" maps. This type comprises the third group spoken of above.

### *(3) Dot Maps*

Upon the basis of the kind of dots used, maps may be divided into three classes. The *first* class is that in which the dots vary in size, each size having a different numerical significance. Such a map is shown in Figure 39. The scale, according to which an illustration is to be drawn, having been

FIGURE 39

PRIMARY MARKETS FOR WISCONSIN CHEESE (AMERICAN) 1911



determined, exact or approximate frequency is indicated in each division of such a map by the number and size of dots. The principle is different from that followed in cross-hatching and coloring. By the use of such dots, actual or approximate frequency is indicated within districts; by the use of cross-hatching and coloring, only group frequency is illustrated. In the former case, each unit of scale may be represented in each district; in the latter case, only one unit is so represented, the complete scale being shown by the entire map. The determining factor in choice of scale, in the first case, is absolute frequency; in the second case, for matter arranged in series, it is the range of the limits of the measures to which the frequencies apply. Grouping is not provided for in the case of the dots and little or no knowledge of geographical distribution is conveyed by exact magnitudes, but only by densities of shades which these magnitudes form. Grouping of frequencies is the cardinal feature of cross-hatched and colored schemes.

As a means of graphically illustrating absolute frequency, such maps are of little value. It is not evident by inspection, and to determine it it is necessary (1) to count the dots, and (2) to evaluate them. In this respect, the method defeats its own end. The process is too tedious and cumbersome. As a method of roughly indicating geographical distribution, they are suggestive, but only with respect to density of shade. In this particular they add nothing to the ordinary cross-hatched type. Moreover, they may give a false impression, two- rather than one-dimensional figures making up the scale of values.[1]

A circle representing a shipment of cheese of 5,000,000 pounds from Wisconsin to Illinois is not easily compared with one representing a shipment of 1,000,000 pounds into Missouri. Again, they are open to the same criticism as cross-hatching *in that they illustrate uniform conditions within and change* only between districts.

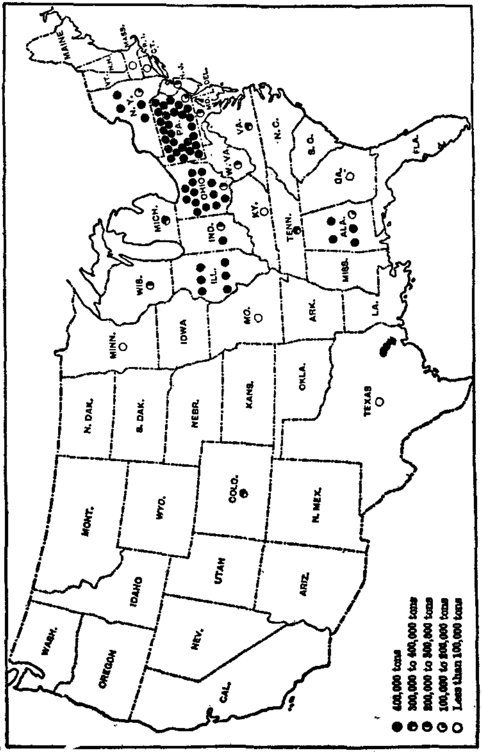
The *second* type of dot maps, as shown in Figure 40, is similar to the first. Instead of using different-sized dots to

[1] The merits of one- and of two-dimensional figures are treated above.

FIGURE 40

Pig Iron Production, by States, 1909

(United States Census, *Statistical Atlas*)



indicate different amounts, uniform sizes are used, the dots being shaded to indicate different values. As a rule the greatest amount is represented by a solid dot, three-quarter, one-half, one-quarter and other shadings indicating lesser amounts. Notwithstanding the fact that such maps are much in vogue, they have little or no advantage over the cross-hatched type. In many respects they are less serviceable.

The *third* type of dot maps, as shown in Figure 41, has certain merits and at the same time certain limitations. The size of the dot is immaterial; the relative frequency with which it occurs is all important. Total frequency is secondary, though in theory it may be approximated, as in the other types of dot maps, by considering the number of dots in connection with the value assigned them. To approximate total frequency is as unnecessary as it is impossible. In most cases the number cannot be determined, because the dots cannot be identified. Moreover, the value assigned to a dot is largely arbitrary, since the purpose of the map is not to record absolute magnitude but to show relative abundance and scarcity in relation to position. The significance of the map is found in the relative *densities* of the dots in different areas. Areas of uniform density are not political jurisdictions, as in colored and cross-hatched maps, but actual positions, so far as the sizes of maps will allow them to be shown.

This form of illustration gives the impression of gradual changes from scarceness to abundance, from "highs" to "lows." It smooths out the breaks which prevail when cross-hatching is used. Geographical barriers are ignored in the drawing, but may be inserted for purposes of study and interpretation. It is easy to visualize places and degrees of concentration and "scatteration"; to get a continuous view of distribution. Dot maps of the third type *suggest* continuous rather than discrete series.

No attempt is made to discuss the technique of diagram and map construction or to enumerate the variety of uses to which diagrams are put by statisticians, publicists, advertisers,

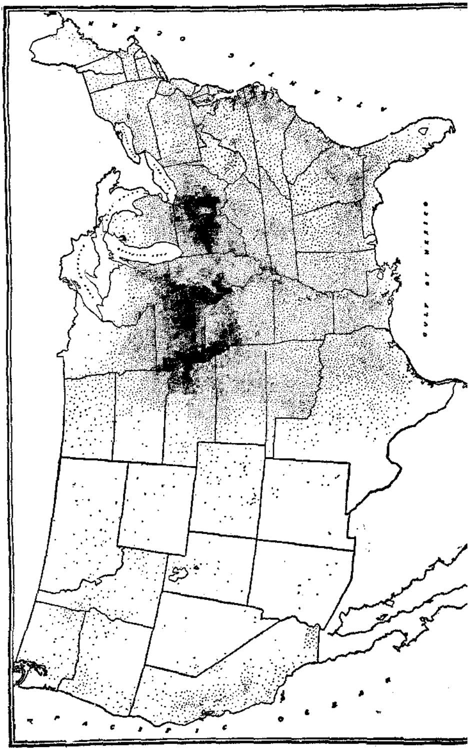


FIGURE 41

NUMBER OF SWINE ON FARMS AND RANGES, APRIL 15, 1910

1 Dot = 2500

manufacturers, etc. Numerous examples of well- and ill-drawn illustrations, together with a discussion of free-hand and mechanical cross-hatching, the uses of pins in map making, preparation of copy for duplicating whether by photographing or otherwise, etc., are given in Brinton: *Graphic Methods for Presenting Facts.*[1] Our interest is more in describing the functions, discovering and defining the limitations of diagrammatic presentation in statistical studies than in describing the processes of drawing and reproducing diagrams, and in indicating their various uses. Such matters are important but they are treated very much more fully elsewhere.

If the reader understands the psychological bases upon which diagrammatic illustration rests—if he appreciates the position which it occupies with respect to tabulation and other steps in statistical analysis, and feels the warning, which it has been the purpose of much of the above to sound, the primary purpose of this discussion will have been realized. The making of diagrams and maps may be left to those who have acquired the requisite skill. The determination to use them should be in the hands of those who have a correct attitude toward their use.

It may be helpful in closing this discussion to outline a few suggestions to be followed in the use of statistical diagrams.

## IV. Suggestions to be Followed in the Use of Statistical Diagrams

(1) Choose illustrations which are least liable to be misunderstood, and which most faithfully and correctly interpret the facts.

(2) See that fact and representation agree, and that all diagrams are provided with concise, clearly stated, and appropriate titles.

(3) Avoid figures which must be read in more than one dimension.

[1] Brinton, Willard C., *Graphic Methods for Presenting Facts*, The Engineering Magazine, New York, 1914.

(4) Indicate on diagrams the scales of values used, and where necessary to avoid confusion, the dimension or dimensions which are significant in interpretation.

(5) Include as a component or as an accompanying part of diagrams the concrete data which they illustrate.

(6) In expressing the different parts of a total, use lines or bars rather than sectors of circles.

(7) In statistical maps representing a series, divide the frequencies and not the number of districts or divisions into equal parts.

(8) In statistical maps representing a series, incorporate as a part of the legend the frequency with which the units of measurement occur, thus indicating the distribution by map and by legend.[1]

### References

BAILEY, W. B., *Modern Social Conditions,* A. C. McClurg, Chicago, 1917, pp. 54-56.

BOWLEY, A. L., *An Elementary Manual of Statistics,* MacDonald & Evans, London, 1915, Chapter V, pp. 35-50.

BRINTON, W. C., "Graphic Methods for Presenting Facts," *Engineering Magazine,* New York, 1914, Chapter I, pp. 1-20; Chapter II, pp. 20-35; Chapter III, pp. 36-53; Chapter IV, pp. 53-69; Chapter XI, pp. 208-227; Chapter XII, pp. 227-254.

HASKELL, ALLAN C., *Graphic Charts in Business,* Codex Book Co., New York, 1922, *passim.*

KARSTEN, KARL G., *Charts and Graphs,* Prentice-Hall, New York, 1923, *passim.*

KING, W. I., *Elements of Statistical Method,* Macmillan & Company, New York, 1912, pp. 91-97.

MARSHALL, WILLIAM C., *Graphical Methods,* McGraw-Hill Book Company, 1921, *passim.*

*Thirteenth Census of the United States,* 1910. Vol. 5, *Agriculture, General Report and Analysis. Statistical Atlas of the United States,* 1910 (1914).

[1] For a further statement of rules to be followed, see "Rules for Diagrammatic Presentation of Statistical Data" in the author's *Readings and Problems in Statistical Methods,* Macmillan & Company, New York, 1920, pp. 273-276.

# CHAPTER VIII

## GRAPHIC PRESENTATION

### I. Introduction

In the preceding chapter, the more common types of diagrams and maps were discussed in their theoretical and practical aspects. It was said that these illustrations—pictograms and cartograms—are subordinate in use to tabulation, coming after it in point of time so far as analysis is concerned, and that they are particularly suited to illustrate statistical series which are discrete or broken. In only one type—the frequency dot map—is there a suggestion of continuity; in the others, whether showing totals or components, the respective parts are distinct.

But there is another type of series which is not discrete nor broken, but continuous. Series of this nature relate to time, to space, and to condition. Time is always continuous, but measurements in time may be continuous or discrete. Temperature measurements at hourly intervals, for instance, are continuous with respect both to the unit (hour) and the measurement (degree). Daily receipts of hogs at Chicago, on the other hand, constitute a series which is continuous as to the unit (day) but discrete as to the number (hogs). The number of farm tractors by counties, for instance, is a space series, continuous as to the unit (county) but discrete as to the measurement (number). A series of words classified according to the numbers of letters which they contain—a condition series—is discrete both as to the unit (number of letters) and the measurements (numbers of instances). So, also, is a series showing the number of hats in a retail in-

ventory, classified by materials from which they are made. On the other hand, the number of people who purchase the hats, a season later, classified according to size of heads, is continuous as to the unit (size) and discrete as to the measurement (number).

Now, it is continuous series with which we are concerned in this chapter. *Diagrams are unsuited to illustrate them.* Other means are necessary if the illustrations are to be true to the facts. Let us see if we can make clear what it is that must be illustrated in such series and the ways in which it may be accomplished.

We shall begin by taking an example of a frequency series, *continuous* as to unit, and discrete as to measurement. An illustration which will suffice for our purpose is the number of employes in a factory, classified by age. The case may be made simple by supposing that an even 100 men were found with ages as follows:

TABLE 22

NUMBER OF EMPLOYES IN FACTORY "X," CLASSIFIED BY AGE

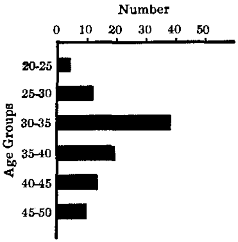
| AGE GROUPS | NUMBER OF EMPLOYES |
|---|---|
| Total | 100 |
| 20 but less than 25 | 4 |
| 25 " " " 30 | 12 |
| 30 " " " 35 | 40 |
| 35 " " " 40 | 20 |
| 40 " " " 45 | 14 |
| 45 " " " 50 | 10 |

The numbers in the different age groups might be shown diagrammatically in a number of ways, but those in Figure 42 are typical.

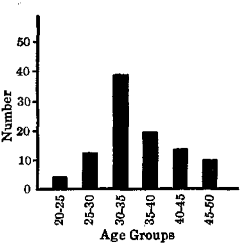
That is, bars indicating the number in each group may be placed horizontally or vertically. These are the conventional diagrammatic types of illustrations.

FIGURE 42

BAR DIAGRAMS SHOWING THE NUMBER OF EMPLOYES IN FACTORY "X" CLASSIFIED BY AGE




*(Readings 20, but less than 25, etc.)*




**(Readings 20, but less than 25, etc.)**

But age is not discrete; it is continuous. The groupings in the illustration are purely arbitrary, and the numbers dependent upon this grouping. Any other groupings—narrower or wider, and starting at any "age"—might be chosen. If other groups are selected, the number in each group will obviously be different. Moreover, the ages as reported, while presumably expressed to the nearest year—"presumably," because of the grouping—are simply approximations to the "true" age—a period susceptible of infinitesimally small gradations. The distinct and separate bars show the ages to be discrete when they are in fact continuous. They should be connected by a continuous line showing that all of the employes fall between the ages, 20± and 50±.

A similar illustration will show the fundamental error in illustrating a continuous time series by a method suitable to one which is discrete. Temperature readings at successive hourly intervals during a day will serve our purpose. Those

chosen are given in Table 23. Diagrammatically, these readings *might* be shown by bars as in Figure 43. But such an illustration is not true to the facts. While the readings vary from hour to hour, neither the temperature nor the time intervals are discrete. Both are continuous. Bars should not be used to illustrate them. The case requires the use of a *line* which will show the gradual and continuous change from one temperature level to another.

TABLE 23

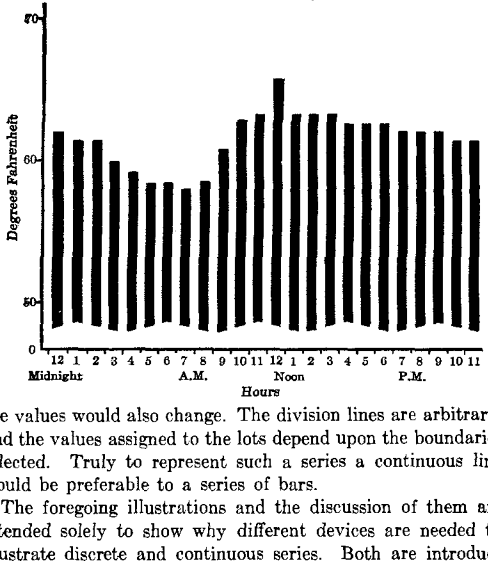
TEMPERATURE MEASUREMENTS AT HOURLY INTERVALS, CHICAGO, SEPTEMBER 3, 1924

| Hours, Sept. 3, 1924 | Temperature — Degrees Fahrenheit | Hours, Sept. 3, 1924 | Temperature — Degrees Fahrenheit |
|---|---|---|---|
| 12 Midnight | 63 | 12 Noon | 68 |
| 1 a.m. | 62 | 1 p.m. | 65 |
| 2 | 62 | 2 | 65 |
| 3 | 60 | 3 | 65 |
| 4 | 59 | 4 | 64 |
| 5 | 58 | 5 | 64 |
| 6 | 58 | 6 | 64 |
| 7 | 57 | 7 | 63 |
| 8 | 58 | 8 | 63 |
| 9 | 61 | 9 | 63 |
| 10 | 64 | 10 | 62 |
| 11 | 65 | 11 | 62 |

An example of a continuous space series may be treated in the same way. Suppose the following data were available showing the value of city property in dollars per front foot for contiguous lots in a city block: Lot 1, $20; lot 2, $15; lot 3, $14; lot 4, $12; lot 5, $14; lot 6, $18; lot 7, $25; and lot 8, $40. Such a series is continuous in fact, although as customarily stated, it is discrete, because of the failure to take account of the *gradual* change from foot to foot. All parts of a given lot are generally assigned the same value. Of course, if the division lines between the lots were changed,

FIGURE 43

BAR DIAGRAM SHOWING HOURLY TEMPERATURE READINGS AT CHICAGO, SEPTEMBER 3, 1924




the values would also change. The division lines are arbitrary, and the values assigned to the lots depend upon the boundaries selected. Truly to represent such a series a continuous line would be preferable to a series of bars.

The foregoing illustrations and the discussion of them are intended solely to show why different devices are needed to illustrate discrete and continuous series. Both are introductory to the more complete discussion of *Graphic Presentation* which follows.

## II. DIAGRAMMATIC AND GRAPHIC PRESENTATION CONTRASTED

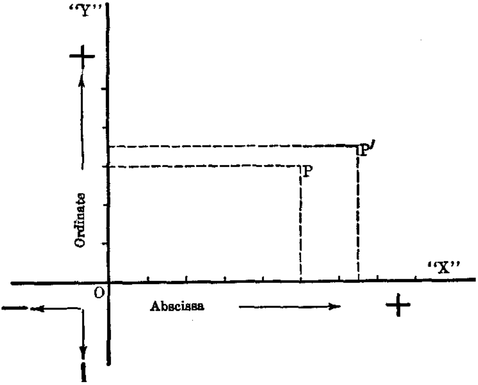
Bars, squares, cubes, circles, and similar figures themselves represent or stand for quantities singly or in series. Such illustrations are diagrammatic. On the other hand, when quantities are *graphically* illustrated, they are not *represented*

by one or more dimensional figures, but are *located* on a surface with respect to two or more dimensions.

The customary method of graphically presenting statistical facts is to use a system of rectangular co-ordinates such as the following:

FIGURE 44

A SYSTEM OF CO-ORDINATES




The points $P$ and $P'$ are two facts located in a plane, their positions being determined by the characteristics indicated on the two axes, $X$ and $Y$. The junction of the axes at $O$ is known as the point of origin; the horizontal axis is called the *abscissa*, and the vertical axis, the *ordinate*. All points in the plane are fixed with reference to these axes. The plus and minus

signs indicate the parts of the "system" in which positive and negative quantities are placed.

Now, it is evident that to *locate* quantities or frequencies *in a plane bounded by two axes in the above fashion* is not the same as to *represent* them by lines, bars, surfaces, solids, etc.—devices which *themselves* are drawn proportional to the amounts or frequencies involved. One would surely not locate quantities or frequencies with respect to these two axes and at the same time represent them by figures of various dimensions. A strange figure, indeed, would be secured if in place of the points $P$ and $P'$ squares or cubes were inserted. If this were done, the co-ordinate axes would have neither use nor meaning. Indeed, it is the function of the ordinate axis to indicate quantities or frequencies, and of the abscissa axis to locate them with respect to time, space, or condition.[1]

In graphic presentation, a system of co-ordinates, such as the above is used; in diagrammatic presentation, the co-ordinates are replaced by the illustrations themselves.

All truly continuous series are properly illustrated by graphical as distinct from diagrammatic methods. Such series, to repeat, may be measured in time or in space or be represented by frequencies of a variable at the same time or place. Since time and frequency series are more commonly encountered in statistical study, they are given primary attention. Let us then begin the study of graphical presentation by considering frequency series.

Time, space, and condition series are contrasted in the chapter on *Classification—Tabulation.*[2] We were there concerned with the manner in which variables in frequency series should be grouped for purpose of tabulation. The problem we found

[1] It is not correct, therefore, to say that "*all statistical diagrams* (?) *are representations of points, lines, surfaces, or solids, the position of which in space are quantitatively defined by a system of co-ordinates.*" Pearl, Raymond, *Introduction to Medical Biometry and Statistics*, W. B. Saunders Company, Philadelphia, 1923, p. 105, italics, the author's.

[2] *Supra*, pp. 157-169.

to be different, depending upon whether such series were discrete or continuous. A similar problem occurs when frequency series are graphically illustrated. It is necessary to know to which type a series belongs before illustrating it.

## III. Graphic Presentation of Frequency Series

### 1. Plotting Simple Frequency Series

Graphically to present statistical facts, two dimensions are used as shown in Figure 44. The horizontal or abscissa axis is used for the measurements, and the vertical or ordinate axis for the frequencies. In any particular case, in order not to over-emphasize the extreme frequencies and at the same time to dwarf the minor ones, it is necessary, before deciding upon the vertical scale, to study the range covered by the frequencies. Similar observations apply to the horizontal axis. If it is divided into units which are too small, the frequencies will be too widely dispersed; if in units which are too large, they will appear crowded. The respective scales will obviously be different for each series of data. There is no absolute standard suitable to all cases, yet, as a general rule, it is desirable to have the horizontal approximately 1½ times as long as the vertical axis. Experience in scale adjustment is the best teacher, however, and a keen sense of form and appearance is helpful while gaining this experience.

*Equal distances on either scale should represent equal facts.*[1] The scales should be divided into units which are easily comprehended in terms of the rulings of the paper used. If paper is ruled in fifths or tenths, for instance, the unit of space on the ordinate should be capable of being readily reduced to this basis. Ten small squares should never be made

[1] On the necessity of having a horizontal as well as a vertical zero base line, see Clark, Earle, "The Horizontal Zero in Frequency Diagrams," in *Quarterly Publications of the American Statistical Association*, June, 1917, pp. 662-669. This article is reprinted in the author's *Readings and Problems in Statistical Methods*, Macmillan & Company, New York, 1920, pp. 385-394.

to equal such an amount as 3,333. A given space should equal some multiple of ten, as 4000, 5000, 6000, etc. The ordinate should be labeled in terms of the scale unit and not in terms of the successive frequencies which are plotted. Exact frequencies may be inserted opposite the measurements to which they apply if they do not crowd the graph. It is well to place them horizontally at the top of the sheet on which the curve is drawn.

The abscissa scale should likewise be divided into equal parts. If for any reason successive units are omitted, given in greater detail, or grouped irregularly, these facts should be plainly indicated by subdividing or widening the unit interval. Under no circumstances should one be left in doubt as to the precise units to which frequencies apply. Uniformity in the size of frequency groups is even more necessary in graphic figures than in tabulation, because an unbroken continuity is more likely to be assumed in the former than in the latter case.

### (1) *Plotting Simple Frequency Distributions of Discrete Series*

The thought was developed above[1] that continuous series cannot properly be illustrated by diagrams. They are designed for those which are discrete. The reverse is equally as true; discrete series cannot properly be illustrated by methods which are suitable to continuous series. Yet, in the case of frequency series which are discrete, continuous lines rather than distinct bars are so commonly (but incorrectly) used that it seems necessary to discuss the problem in detail.

Measurements in discrete series, by custom or otherwise, are expressed in the units in which the thing measured is shown. Many illustrations of such series have already been given. When they are graphically presented, the units on the abscissa axis do not represent approximations to exact measurements which it is impossible to determine because of

[1] Pp. 215-218.

the limitations of science, or because all possible measurements are likely to occur within the limits set up. They represent actual measurements. The units on the abscissa axis assigned to them, therefore, can rarely be accurately represented by spaces. They are almost always points or positions.

Moreover, if lines are used to connect the ordinates, they are meaningless. It is true that they aid the eye in comparing the respective heights of the ordinates, but beyond this they serve no purpose. They show a trend of frequencies at the positions at which they occur but they *do not* indicate the likely or probable frequencies at every point on the horizontal axis, as would be the case with a line describing a continuous series.

This can be made clear by means of examples. A recent study showed that certain proposed freight rates per one hundred pounds from St. Paul and Minneapolis to Sioux City, Iowa, were expressed in amounts ending in integers as follows:

TABLE 24

Proposed Freight Rates per 100 Pounds Between St. Paul, Minneapolis, and Sioux City, Iowa, Ending in Different Integers

| Integers | Number of Rates |
|---|---|
| 0 | 6 |
| 1 | 3 |
| 2 | 7 |
| 3 | 4 |
| 4 | 6 |
| 5 | 4 |
| 6 | 6 |
| 7 | 4 |
| 8 | 6 |
| 9 | 4 |

Suppose this frequency series were graphically illustrated by a *continuous line* running from zero to nine, inclusive. It

would then appear that something more than four and something less than six cases, for instance, occurred between amounts ending with integers of three and four. But such an inference would be absurd. There are no integers between three and four. Accordingly, separate bars, rather than a continuous line, should be used to illustrate this series.[1]

Table 25 on page 226 shows the number of employes in mercantile establishments classified according to rates of wages received. This, obviously, is a discrete series. While weekly wage-rates other than those actually named might have been paid, there is no basis for assuming that the difference in frequencies between 254 and 4, for \$6.00 and \$6.50, respectively, are evenly distributed between these two amounts, or that there are any persons whatsoever who receive \$6.399, for instance. A continuous line connecting the different ordinates in such a case as this may serve to emphasize the difference, but it does not establish the distribution between them.

It is customary in illustrating discrete series to represent group-widths by spaces on the abscissa axis, to erect ordinates at their middle points, and to connect them by continuous lines. This practice is bad, because it makes it appear that there is either an equal distribution of the instances throughout the groups, or that they are all concentrated at their centers. In most cases, neither condition obtains. There is no necessity that such a distribution should hold for a discrete series.[2] Indeed, any grouping at all for such series is

[1] In an analogous case, *The Bureau of Railway Economics*, in plotting the "Monthly Revenues and Expenses per Mile of Line" for the railroads in the United States having operating revenues above \$1,000,000, says, "The points on the vertical lines are of significance only in showing the condition for the particular month. The lines connecting the points assist in tracing the change from month to month but do not indicate the trend during the month, nor do they represent cumulative figures for the period." "Revenues and Expenses of Steam Roads in the United States, December, 1915," *Bureau of Railway Economics*, Washington, D. C.

[2] It is known, for instance, that wage-rates are generally fixed in round numbers, concentration appearing on 5, and its multiples. See Table 25, and the following distribution.

likely to be misleading. If possible, each measurement should be separately indicated. This, of course, is impossible in many cases: some grouping must be used. A graphic figure, however, should, so far as possible, faithfully represent the facts as they are. It should never imply a distribution which does not exist. If it is an error to connect by *straight* lines ordinates representing frequencies in discrete series—because of implications as to distribution—it is a far greater error to connect them by *smoothed* lines. If series are discrete, it is this very characteristic which should be retained: false accuracy is implied when a smoothed line is used. Only when such a line gives an accurate notion of direction at, and change between successive measures should it be used. It should not be employed as a means of generalizing as to distributions at measures not represented.

It is doubtful if the distribution of interest rates on real estate mortgages, for instance, as shown in Chapter V,[1] would have been materially altered by extending the study over a longer period of time, or by including more instances. Smoothing such curves results in deception. *Smoothing may be employed to remove errors in observation but not to disguise the truth.* The extent to which it does the latter varies directly,

(Note 2, continued)

Table showing the number of union bricklayers receiving specified hourly wage-rates in New York State. (Compiled from the New York Department of Labor Bulletin, Whole No. 65, 1913, pp. 4-6)

| Cents per Hour | Number | Per Cent Distribution |
|---|---|---|
| Total | 13,362 | 100.00 |
| 50 | 496 | 3.71 |
| 55 | 489 | 3.66 |
| 60 | 1,650 | 12.35 |
| 65 | 2,391 | 17.89 |
| 70 | 7,404 | 55.42 |
| All other | 932 | 6.97 |

[1] Page 164.

TABLE 25

TABLE SHOWING THE NUMBER OF FEMALES AND MINORS EMPLOYED IN 24 MERCANTILE ESTABLISHMENTS IN SEPTEMBER, 1913, RECEIVING CLASSIFIED WAGE-RATES

("Minimum Wage Legislation in the United States and Foreign Countries"—*Bulletin of the United States Bureau of Labor Statistics*—Whole Number 167, April, 1915, p. 96)

| Weekly Wage-Rates | Number of Females and Minors Receiving Specified Wages | Weekly Wage-Rates | Number of Females and Minors Receiving Specified Wages |
|---|---|---|---|
| Total | 3,189 | | |
| $3.00 | 20 | $14.00 | 60 |
| 3.50 | — | 14.50 | 2 |
| 4.00 | 50 | 15.00 | 164 * |
| 4.50 | 18 | 15.50 | 2 |
| 5.00 | 72 | 16.00 | 27 * |
| 5.50 | 2 | 16.50 | 15 |
| 6.00 | 254 * | 17.00 | 14 |
| 6.50 | 4 | 17.50 | 26 |
| 7.00 | 311 * | 18.00 | 65 * |
| 7.50 | 48 | 18.50 | 4 |
| 8.00 | 490 * | 19.00 | 5 |
| 8.50 | 44 | 19.50 | 4 |
| 9.00 | 441 * | 20.00 | 57 * |
| 9.50 | 4 | — | — |
| 10.00 | 370 * | 21.00 | 3 |
| 10.50 | 13 | 22.00 | 23 |
| 11.00 | 72 * | — | — |
| 11.50 | 8 | 25.00 | 37 * |
| 12.00 | 355 * | 27.50 | 7 |
| 12.50 | 16 | 30.00 | 9 |
| 13.00 | 22 | — | — |
| 13.50 | 37 | 35.00 | 9 |
| | | Over 35.00 | 5 |

* **Notice the concentration on even dollar amounts.**

for discrete series, with the degree of irregularity characteristic of the thing measured and with the widths of the groups into which frequencies are placed.

This discussion, however, is in fact out of place. Discrete series of the frequency type should be illustrated by diagrams —discrete figures. The subject is discussed in this chapter only because this fact is so often forgotten or ignored. Continuous lines—straight and smoothed—and bar diagrams are used indiscriminately to illustrate both continuous and discrete series. Both principle and consistency are sadly lacking in these respects. But they ought not to be.

### (2) *Plotting Simple Frequency Distributions Describing Continuous Series*

When plotting continuous frequency series, the case is different. The units of measurement are arbitrary, the frequencies being functions of those selected. Accordingly, the abscissa axis, properly considered, is continuous. The breaks in it are made for convenience only: they indicate conventionally "stops," as it were. They are artificial. If this is so, then, the ordinates indicating the frequencies at these "stops" should be connected by smooth lines which suggest continuity in the thing measured. To regard the measurements actually made as fully descriptive of such a series, is as incorrect as it is to assume, in the case of discrete series, that instances occur at all possible measurements. Neither is correct. One type of illustration fits a continuous, the other a discrete, series.

In continuous series, since variations from one extreme measurement to the other are regular and gradual, not only should the ordinates be connected, but the direction of the line joining them should be determined by the frequencies at successive and at all measures. Such a curve should be free from sharp angles, the contour being influenced at each point by the relative sizes of adjoining frequencies and by the character of the complete distribution.

Let us take a continuous frequency series and see how it would be correctly illustrated graphically. For this purpose the measurements of the lengths of 327 ears of corn taken at random from a homogeneous "population" may be selected.[1] Measurements are made to the nearest quarter of an inch and grouped into one-half inch classes. The following table shows the number of ears falling into the half-inch groups.

TABLE 26

TABLE SHOWING THE NUMBER OF EARS OF CORN CLASSIFIED BY LENGTHS

| LENGTH OF EARS OF CORN IN INCHES | NUMBER OF EARS AT EACH LENGTH |
|---|---|
| Total | 327 |
| 3.0 | 1 |
| 3.5 | 0 |
| 4.0 | 1 |
| 4.5 | 0 |
| 5.0 | 2 |
| 5.5 | 3 |
| 6.0 | 9 |
| 6.5 | 8 |
| 7.0 | 12 |
| 7.5 | 19 |
| 8.0 | 32 |
| 8.5 | 40 |
| 9.0 | 67 |
| 9.5 | 63 |
| 10.0 | 38 |
| 10.5 | 21 |
| 11.0 | 8 |
| 11.5 | 2 |
| 12.0 | 1 |

The precision of the measurements and the widths of the groups determine the number of ears in each class. If the

[1] Data taken from Davenport, Eugene, and Rietz, Henry L., "Type and Variability in Corn," *Bulletin 119, University of Illinois Agricultural Experiment Station*, October, 1907, p. 3.

measurements had been made to the nearest tenth of an inch and grouped into quarter-inch classes, as 4.00, 4.25, 4.50, 4.75, 5.00, 5.25, 5.50, 5.75, 6.00, 6.25, etc., then "at 5.75 would be grouped all ears which measured 5.7 and 5.8, while at 5.00 would be grouped those which measured 4.9, 5.0, and 5.1. In the long run, this would clearly result in placing more ears at 5.0 than at 5.25, other things being equal. If we should group measurements taken to the nearest tenth inch in 0.5 inch or 0.3 inch classes, no such difficulty arises."[1]

With the grouping shown in Table 26, it is absurd to assume that since 40 ears are grouped at 8.5 inches, and 67 at 9.0 *in length, there were no ears with lengths between these* measurements. Had they been more precise and the groupings narrower—thus giving a different distribution from that shown in the table—each measurement would still have been an approximation to the "true" length, and the grouping arbitrary. The unit of measurement is strictly continuous—any break in it is artificial.

But the ears measured are only a sample of a wider "universe." Would the case be different if more cases were taken? Not at all. There would still be the problem of determining the length of each ear, and for this purpose an approximation—no matter how precise—would have to be made. Length is continuous, and merely increasing the number of cases in which it must be determined does not alter the fact that each measurement of length is an approximation.

In order to illustrate graphically the number of ears at each of the lengths shown in the groups in Table 26, a continuous, smooth line from ordinate to ordinate should be used. The case in this respect would be no different if the sample were enlarged.

The degree to which continuous frequency series may be smoothed depends upon the nature of the distributions. If measurements are accurately made—bias due to personal and

[1] *Op. cit.*, p. 28.

mechanical elements being absent or distributed according to chance—large deviations from a standard will be less common than small ones, the measurements tending to be arranged around an average or norm. This is the case with distributions approaching the "normal law of error" type.[1] According to this "law," the measurements of phenomena are distributed about their averages in a regular and systematic manner, when the number observed is large, and when each measurement results from a large number of independent causes, none of which is of preponderating importance. A graphic figure of such a distribution is bell-shaped in form, the precise form being dependent upon the degree to which chance operates, and upon the number of measurements made.

The measurements of the lengths of a sufficient number of ears of corn would tend to give such a distribution. Indeed, it tends to be characteristic of the measurements of all natural phenomena. Accordingly, in smoothing distributions of this type, account should be taken of the tendency for frequencies, as they approach the maximum ordinate or most common measurement, to pile up at the upper sides, and as they recede from the maximum, to pile up at the lower sides, of the groups into which they are placed. Allowance should be made for this tendency in smoothing the distributions of the measurements of a sample, as well as in generalizing as to the distribution of an entire "population."

In the illustration of the lengths of ears selected, 240 cases occur in the groups 8.0 to 10.0 inches. The greatest number—67—is found at 9.0 inches. At the one-half inch below, there are 40, and at the one-half inch above, 63 cases. That is, the distribution is unequally balanced near the maximum and "tails off" more below than above it. It is not strictly of the normal type. If more ears were included in the sample, the form of distribution would appear more regular. Accordingly, in smoothing the curve to take account of

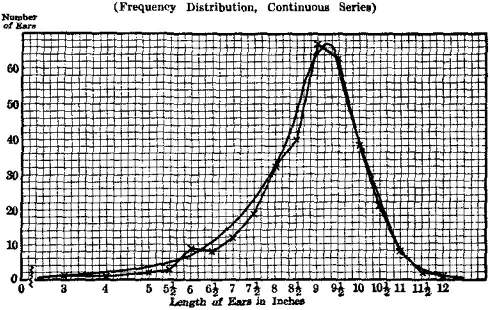
[1] See Chapter XI, pp. 367-370.

this fact, a continuous line should be drawn near to but not at the various points in the distribution. The curve used to smooth the sample measurements should be rounded out as the larger frequencies are approached and inclined toward the vertical as they fall off. The smooth curve in Figure 45 is intended to "fit" the sample and not to generalize the distribution of an *ideal* curve relating to such measurements.

FIGURE 45

SMOOTHED FREQUENCY DISTRIBUTION OF LENGTHS OF EARS OF CORN

(Frequency Distribution, Continuous Series)




In any continuous series, as the class intervals into which measurements are grouped are made smaller, and as the accuracy of measurement is made more precise—the number of observations being large—the lines drawn from successive ordinates appear smooth and regular. On the other hand, if the observations are few, or, if the groups into which they are placed are chosen without regard to the distribution in normal curves, then the lines connecting successive ordinates have a step-like, halting appearance, foreign to continuous series. In grouping data of the continuous type, the "classes should be

only just broad enough to make the distribution fairly smooth, that is, there should be no vacant classes except very near the extremes of the range, and a gradual increase from one extreme up to the maximum and then a gradual decrease to the other extreme, if there is only one maximum in the distribution as is, in general, the case with these populations."[1] A smoothed curve serves the purpose of idealizing such a grouping in keeping with the normal type of distribution. Any pronounced tendency of distribution in a continuous series, shown by a fair and adequate number of samples, will tend to be confirmed if more are taken. On the other hand, if only a few are studied and the resulting curve tends to be very irregular, it is likely that further sampling will give a more characteristic tone to the distribution, making less pronounced both the exceptionally large and small frequencies. Whether a smoothed curve should exaggerate or minimize the peculiar properties of a distribution depends upon how accurately the samples characterize the complete series.[2]

How fully this is done by any series of samples is not always evident. While some smoothing is always admissible for continuous series, smoothed curves should not be used indiscriminately in place of the original data. The measurements of the samples and the frequencies with which they occur often serve as the best available approximation to the ideal which it is the purpose of the smoothed curve to give.

## 2. PLOTTING CUMULATIVE FREQUENCY SERIES

The foregoing discussion of graphic representation has had to do with *simple* frequency series: that is, series in which the

[1] Davenport, Eugene, and Rietz, Henry L., *op. cit.*, p. 27.

[2] *To the rule* "that the top of the curve usually overtops the highest point of the frequency polygon, especially when the classes are rather large" (King, W. I., *Elements of Statistical Method*, Macmillan & Company, New York, 1912, p. 113), the criticism is pertinent that the determining factor is not so much the size of the groups as it is the representative character of the samples.

numbers of instances refer to the respective measurements or to the groups into which they are placed. But the frequencies may be cumulated: that is, added together, the effect of this being to include together successive measurements or groups as the case may be. Each frequency class, therefore, is made to include all of the lower or all of the upper classes, depending upon the manner in which the cumulating is done. It may be begun with either extreme measurement, the only essential being, if all cases are to be included, that it be carried through the entire range of frequencies. If it proceeds from the least to the greatest, the frequencies at each step are read "less than"; if from the greatest to the least, "more than." It will be noticed that the cumulations when read "less than" refer to the upper limits, and when read "more than," to the lower limits of the respective groups. This method of stating the frequencies is used in Table 29.

Both discrete and continuous series may be cumulated and the resulting frequencies graphically illustrated. The way in which the cumulating is done is the same in both series but the graphic representations are different. The following discussion will serve to make this clear.

### (1) "Graphic" Representation of Discrete Frequency Series Cumulated

The discrete series in Table 25, p. 226, may be cumulated on a "less than" or on a "more than" basis. Within the limits set by the simple series, any grouping desired may be used. Different methods of cumulating are shown in Tables 27 and 28.

A system of rectangular co-ordinates, as shown in Figure 44, is used to illustrate cumulative as well as simple frequency distributions. The groups are measured on the abscissa or $X$ axis, and the frequencies, on the ordinate or $Y$ axis, equal distances on either axis always representing equal quantities as in the case of simple frequency series. When the successive groups are indicated from left to right along the $X$ axis, the frequencies cumulated on a "less than" basis tend to in-

crease, successive intervals including all of the frequencies which belong to the lower classes as well as those at a given position. When they are cumulated on a "more than" basis, the frequencies from left to right tend to decrease, successive intervals including only the remaining frequencies as well as those in the class in question.

TABLE 27

CUMULATIONS OF WEEKLY WAGE-RATES ON A "LESS THAN" BASIS
(Simple Frequency Series, p. 226)

"A"

| Weekly Wage-Rate Groups | Cumulated Frequencies |
|---|---|
| Total | 3189 |
| Less than $ 5.00 | 88 |
| " " 10.00 | 1758 |
| " " 15.00 | 2713 |
| " " 20.00 | 3039 |
| " " 25.00 | 3122 |
| " " 30.00 | 3166 |
| " " 35.00 | 3175 |
| " " 40.00 * | 3189 |

"B"

| Weekly Wage-Rate Groups | Cumulated Frequencies |
|---|---|
| Total | 3189 |
| Less than $ 8.00 | 779 |
| " " 16.00 | 2879 |
| " " 24.00 | 3122 |
| " " 32.00 | 3175 |
| " " 40.00 * | 3189 |

* Limit arbitrarily taken.

TABLE 28

CUMULATIONS OF WEEKLY WAGE-RATES ON A "MORE THAN" BASIS
(Simple Frequency Series, Table 25, p. 226)

"A"

| Weekly Wage-Rate Groups | Cumulated Frequencies |
|---|---|
| Total | 3189 |
| More than $20.00 | 93 |
| " " 15.00 | 312 |
| " " 10.00 | 1061 |
| " " 5.00 | 3029 |
| " " 0.00 | 3189 |

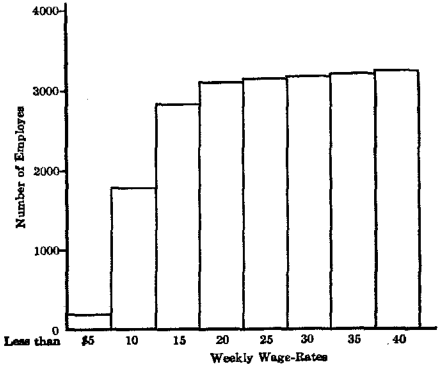
"B"

| Weekly Wage-Rate Groups | Cumulated Frequencies |
|---|---|
| Total | 3189 |
| More than $22.00 | 67 |
| " " 18.00 | 163 |
| " " 14.00 | 478 |
| " " 10.00 | 1061 |
| " " 6.00 | 2773 |
| " " 0.00 | 3189 |

To combine the frequencies by successively widening the groups does not change the fundamental nature of truly discrete series. The frequencies, whether expressed in simple or cumulated form, are distinct at each measure encountered. Accordingly, a continuous line, whether irregular or smoothed, ought not to be used to illustrate them. Successive accumulations should be indicated by separate bars located at the abscissa units. For instance, the cumulations on a "less than" basis, as shown in part "A" of Table 27, would appear as in Figure 46.

FIGURE 46

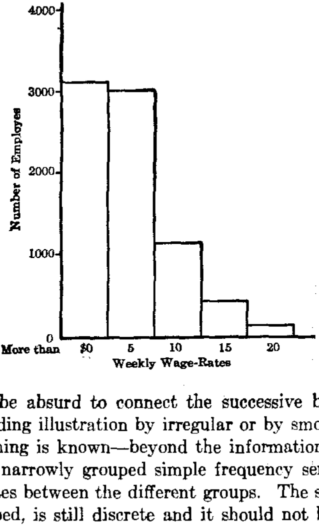
BAR DIAGRAM SHOWING A DISCRETE FREQUENCY SERIES CUMULATED ON A "LESS THAN" BASIS




On the other hand, if the series as cumulated in part "A" in Table 28—that is, on a "more than" basis—were illustrated, the figure would appear as in Figure 47.

FIGURE 47

Bar Diagram Showing a Discrete Frequency Series Cumulated on a "More Than" Basis




It would be absurd to connect the successive bars in this or the preceding illustration by irregular or by smoothed lines because nothing is known—beyond the information contained in the more narrowly grouped simple frequency series—about the wage-rates between the different groups. The series, howsoever grouped, is still discrete and it should not be made to appear continuous.

If cumulations were made at precise amounts, as, for instance, those in Table 25, the successive ordinates should be drawn at intervals so marked on the abscissa axis. Moreover, they should not be connected in any way. The amounts are discrete and they should be so represented.

So much for the representation of discrete series. In what way is graphic illustration different in the case of series which

are continuous? This question is answered in the following section.

### *(2) Graphic Representation of Continuous Frequency Series Cumulated*

Frequency series may be continuous as to the unit of measurement and discrete as to the frequencies. Let us take an example of such a series and discuss its graphic representation when the instances are cumulated.

Table 29 shows the number of towns in the United States classified according to the prices paid for oil in 1904. The unit (price) is in fact continuous, although as customarily stated it is discrete. In this case, we shall consider it to be continuous. For purposes of illustration, one-tenth part of a cent is taken as a convenient, although arbitrary, division. The frequencies, however, are discrete, numbers of instances being used.

The second column of Table 29 shows a simple frequency distribution of the towns classified according to prices paid. Columns three and four, respectively, show the frequencies cumulated on a "less than" and on a "more than" basis. Cumulative graphs or ogives of the series are shown in Figure 48. The direction of the "less than" curve is from the lower left-hand to the upper right-hand corner; and that of the "more than," from the upper left- to the lower right-hand corner of the figure.

As the cumulations are made in Table 29, and as they should be read on the curve, the frequencies which are expressed on a "less than" basis always refer to the upper sides, and those on a "more than" basis to the lower sides of the groups. For instance, the number of towns where prices are 10 cents or less is 914; the number, in which they are more than 10 cents is 91C.

In graphically illustrating this series, the respective ordinates showing the number of towns are connected by straight

TABLE 29

TABLE SHOWING THE DISTRIBUTION OF TOWNS ACCORDING TO PRICES PAID FOR OIL, FREIGHT DEDUCTED (1830 QUOTATIONS), DECEMBER, 1904, FOR THE UNITED STATES

(*Report of the Commissioner of Corporations on the Petroleum Industry*, Part II, Aug. 5, 1907, p. 951)

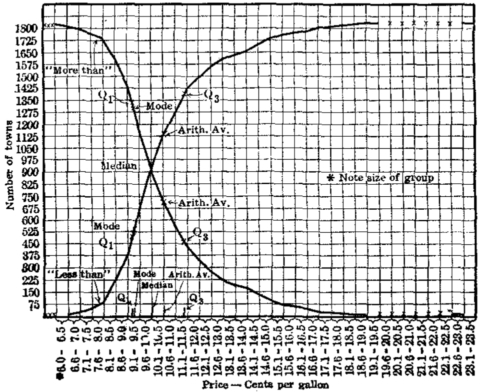
| PRICE, LESS FREIGHT (Cents per gallon) | NUMBER OF TOWNS IN THE UNITED STATES | | |
|---|---|---|---|
| | Simple Frequency | Cumulative Frequency | |
| | | "Less than" | "More than" |
| Total | 1,830 | — | — |
| 6.0 to aud including 6.5 | 11 | 11 | 1,830 |
| 6.6 to and including 7.0 | 17 | 28 | 1,819 |
| 7.1 to and including 7.5 | 27 | 55 | 1,802 |
| 7.6 to and including 8.0 | 36 | 91 | 1,775 |
| 8.1 to and including 8.5 | 123 | 214 | 1,739 |
| 8.6 to and including 9.0 | 181 | 395 | 1,616 |
| 9.1 to and including 9.5 | 281 | 676 | 1,435 |
| 9.6 to and including 10.0 | 238 | 914 | 1,154 |
| 10.1 to and including 10.5 | 201 | 1,115 | 916 |
| 10.6 to and including 11.0 | 162 | 1,277 | 715 |
| 11.1 to and including 11.5 | 130 | 1,407 | 553 |
| 11.6 to and including 12.0 | 85 | 1,492 | 423 |
| 12.1 to and including 12.5 | 65 | 1,557 | 338 |
| 12.6 to and including 13.0 | 49 | 1,606 | 275 |
| 13.1 to and including 13.5 | 26 | 1,632 | 224 |
| 13.6 to and including 14.0 | 19 | 1,651 | 198 |
| 14.1 to and including 14.5 | 43 | 1,694 | 179 |
| 14.6 to and including 15.0 | 38 | 1,732 | 136 |
| 15.1 to and including 15.5 | 23 | 1,755 | 98 |
| 15.6 to and including 16.0 | 12 | 1,767 | 75 |
| 16.1 to and including 16.5 | 13 | 1,780 | 63 |
| 16.6 to and including 17.0 | 20 | 1,800 | 50 |
| 17.1 to and including 17.5 | 8 | 1,808 | 30 |
| 17.6 to and including 18.0 | 7 | 1,815 | 22 |
| 18.1 to and including 18.5 | 6 | 1,821 | 15 |
| 18.6 to and including 19.0 | 4 | 1,825 | 9 |
| 19.1 to and including 19.5 | 1 | 1,826 | 5 |

TABLE 29—*Continued*

| Price, Less Freight (Cents per gallon) | Number of Towns in the United States | | |
|---|---|---|---|
| | Simple Frequency | Cumulative Frequency | |
| | | "Less than" | "More than" |
| 19.6 to and including 20.0..... | — | — | — |
| 20.1 to and including 20.5..... | — | — | — |
| 20.6 to and including 21.0..... | — | — | — |
| 21.1 to and including 21.5..... | — | — | — |
| 21.6 to and including 22.0..... | — | — | — |
| 22.1 to and including 22.5..... | — | — | — |
| 22.6 to and including 23.0..... | 1 | 1,827 | 4 |
| 23.1 to and including 23.5..... | 3 | 1,830 | 3 |

FIGURE 48

Cumulative Graphs—Ogives—Constructed on "More Than" and "Less Than" Bases, Showing by Towns the Classified Prices of Oil



but irregular lines. This is admissible because from ordinate to ordinate the price differences are gradual, each amount being only an approximation (in this instance to the nearest tenth of a cent) to the "true" price. Such lines represent the gradual changes, but they do not idealize them as would a smoothed curve, designed to "fit" the distribution. The continuous lines are intended to illustrate the measurements in this particular sample, rather than to generalize from it as to the nature of the distribution from a total "population" of this sort.

The frequencies in a continuous series may be indicated as relating to a precise measurement. This is done in the example showing the number of ears of corn of different lengths.

Each measurement, as was said, is only an approximation to the "true" length. The case is the same in the following example showing the lengths of time in 61 trials which it takes a mechanic to "thread" a standard bolt. The number of fre-

TABLE 30

LENGTHS OF TIME TAKEN TO "THREAD" A STANDARD BOLT
(Measurement to nearest quarter of a minute)

| MINUTES | FREQUENCIES | |
|---|---|---|
| | Simple | Cumulated "Less than" |
| Total | 61 | — |
| 5¼ | 2 | |
| 5½ | 3 | 5 |
| 5¾ | 5 | 10 |
| 6 | 6 | 16 |
| 6¼ | 8 | 24 |
| 6½ | 12 | 36 |
| 6¾ | 9 | 45 |
| 7 | 7 | 52 |
| 7¼ | 4 | 56 |
| 7½ | 3 | 59 |
| 7¾ | 2 | 61 |

quencies at each time—approximations to the nearest quarter of a minute—are given in Table 30.

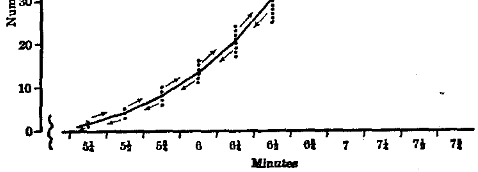
Suppose it were desired graphically to illustrate the *cumulated frequencies at successive intervals.* Since time is continuous, the frequencies in reality have reference not to the measurements as stated but to approximations to them. Accordingly, account should be taken of this fact in the graphic figure. The way in which it is done is illustrated in Figure 49, and may be described as follows:

At each successive interval on the abscissa axis, the number of frequencies is indicated by dots according to the scale provided on the ordinate. Beginning at the shortest time, 5¼ minutes, two dots equally spaced are inserted. With these as the total for this period, three dots for the second interval, 5½ minutes, are added with the upper dot for the preceding time unit as a base. This process is continued until the fre-

FIGURE 49

CUMULATIVE GRAPH OF A CONTINUOUS FREQUENCY SERIES SHOWING LENGTH OF TIME TAKEN TO "THREAD" A STANDARD BOLT

(Basis of Cumulation—"Less Than")



quencies at the different positions are inserted. A continuous line is then drawn through the middle points of the consecutive vertical rows of dots. It is this line which properly represents the cumulations. This follows because (1) not all of the frequencies assigned to the respective measurements actually fall upon them—they fall "around" them, and (2) there are probably as many measurements in excess as in defect of the approximate time, the number of instances being uniformly distributed over a quarter of a minute. Accordingly, if the dots, each of which represents an approximate period, are supposed to lie upon the continuous line rather than to have a vertical position, the continuity of the series is illustrated. The positions which they would then assume are indicated on the figure by the small arrows.

Continuous straight lines connecting the middle points of the different ordinates properly illustrate the nature of the cumulation in this *sample*. If, however, it were taken to characterize a "population" of this sort, the connecting line should be smooth and free from all angles.

From the foregoing discussion, it should be apparent that the methods of graphically illustrating simple and cumulated discrete and continuous frequency series are fundamentally different. Choice of methods depends upon the nature of the series. No careful student will select methods purely at random. The requirements in each case are different and these should be observed. Graphic figures should not only be accurately drawn but selected according to their appropriateness. To make such selection calls for more than mere cleverness and ability to draw.

## IV. Graphic Presentation of Historical or Time Series[1]

The ways in which discrete time series should be illustrated are discussed in Chapter VII under the heading *Diagrammatic*

[1] See also Chapter XIV, *passim*.

*Presentation.* Time, of course, is continuous, but as has been said in a number of places, measurements in time may be discrete or continuous. Those of the first type should be indicated by vertical or horizontal bars; those of the second, by unbroken lines.

In graphically presenting continuous time series, a number of problems present themselves. These have to do with (1) choice and adjustment of scales, (2) the type of lines connecting successive ordinates, and (3) curve smoothing. In keeping with the outline plan of treatment of frequency series, *simple* and *cumulative historical curves or graphs* will be discussed separately.

## 1. PLOTTING SIMPLE HISTORICAL SERIES

*Simple* historical series are those in which amounts or frequencies relate to instants or intervals of time. *Cumulated* historical series, on the other hand, are those in which amounts or frequencies are totaled at successive instants or intervals of time. It is the first type with which we are now concerned.

### *(1) Choice and Adjustment of Scales*

A system of rectangular co-ordinates, as shown in Figure 44, is used to illustrate time series. The time units are placed on the abscissa or $X$ axis, and the amounts or frequencies on the ordinate or $Y$ axis. Since time has no beginning, a horizontal zero is unnecessary; the first units may, as convenience demands, be indicated near or removed from the point of origin at the intersection of the two axes. The ways in which the time units are shown, however, differ according to the nature of the measurements. If they are taken at successive instants, as would be the case, for example, in the measurements of temperature, the *unit* on the horizontal axis is indicated as a *point*. If the measurements are in the nature of totals which accumulate *during* a period, as would be the case,

for instance, in sales by years, then the unit on the abscissa is indicated as a *space*. In both cases, however, the abscissa axis should be divided into equal parts, each one representing instants equally removed or periods of equal length.

a. Natural Scale or "Difference" Charts

The ordinate scale, when amounts or frequencies are shown, should begin with zero, since they are always reckoned from it as a starting point. If this rule cannot be followed, attention to its violation should be indicated in some unmistakable way. This can be done by using a star (*) and a footnote calling attention to the fact, or better by drawing a wavy (~~~) line across the ordinate axis and parallel to the $X$ axis. Equal space units on the ordinate scale should represent equal amounts. But "equal amounts" may have reference to quantities or to ratios, and these are not the same. *If a scale of ratios is used, a zero line is unnecessary*—in fact, there is no zero in such cases.[1]

In deciding upon the proportions between the respective scales, the aim should be (1) to allow ample room for the illustration itself and for the data which it shows to be included on the graph, (2) neither to over-emphasize nor to dwarf the extreme fluctuations, (3) to bring out the characteristics of the changes over the entire period and from time to time (instant or interval). Bowley states the problem and the way in which it should be met in the following language:

"It is only the ratio between the horizontal and the vertical scales that needs to be considered. The figure must be sufficiently small for the whole of it to be visible at once; if the figure is complicated, relating to a long series of years and varying numbers, minute accuracy must be sacrificed to this consideration. Supposing the horizontal scale decided, the vertical scale must be chosen so that the part of the line which shows the greatest rate of increase is well inclined to the vertical, which can be managed by making the scale sufficiently small; and, on the other hand, all important fluctuations

[1] See the discussion of *Ratio Scales and Ratio Charts*, *infra*, pp. 248-255.

must be clearly visible, for which the scale may need to be increased. Any scale which satisfies both of these conditions will fulfill its purpose."[1]

The two scales selected will, in a given case, depend, among other things, upon the size of the page, the ability of the eye to view the illustration as a whole, and the subsequent uses to which it is to be put. In the latter respect, a graph used as a working paper will differ from one prepared for publication. The above discussion of the proportions between the respective scales has reference to illustrations involving but a single series. When two or more curves are to be placed in the same illustration, the case is complicated in the following, among other, ways: (1) the amplitude of the fluctuations may be noticeably different, (2) they may refer to different periods of time, (3) they may be measured in units of widely different size, or in entirely different units. Any or all of these conditions necessitate compromises of one sort or another to be made.

If the amplitudes of the fluctuations are widely different, and one chart is used, *two* ordinate scales may be required if *actual* amounts are plotted—that is, if equal spaces show equal amounts rather than equal ratios. The same may be true if the amounts differ greatly in size. In this case, a single scale may be used if it is broken or made discontinuous, one portion fitting the smaller, and one the larger amounts. The place at which the scale is broken should be indicated by a wavy line (〰〰) being drawn across the entire chart. To do this has the advantage of bringing the two parts of the charts closely together, but the disadvantage of leaving the upper part without an *evident* zero base. This is to be avoided whenever possible. In such cases, it is preferable to use separate scales, both beginning at zero.

When two or more series of data are placed on a single chart, it is sometimes necessary, when difference rather than

[1] Bowley, A. L., *Elements of Statistics*, King, London, 1911, p. 149.

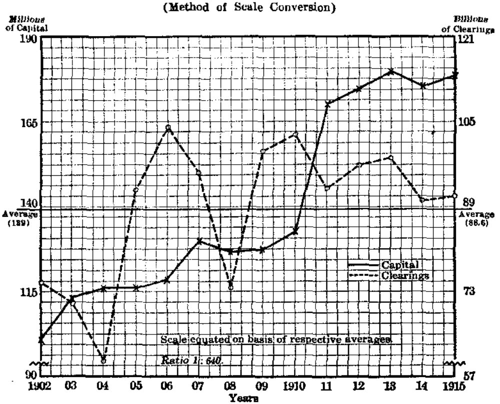
ratio changes are shown, to convert one scale into terms of the others. Some of the ways in which this can be done are as follows:

(1) By choosing separate scales and making each proportional to the respective averages of the series. Such an adjustment is made in Figure 50. Each of the curves must then be read in terms of its own scale—the amounts being in fact deviations, plus and minus, from their respective averages.

FIGURE 50

CAPITAL AND CLEARINGS OF NEW YORK CLEARING HOUSE BANKS, 1902-1915

(Method of Scale Conversion)




(2) By expressing the items in the series as percentages of their respective totals, and plotting the deviations. When two or more series treated in this form are plotted on a single chart (*a*) relative rather than absolute quantities are shown, and (*b*) the respective curves may be far removed from each

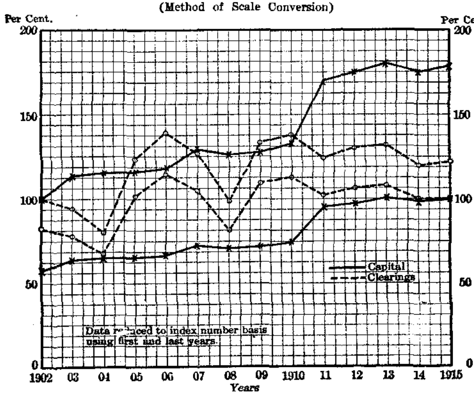
other, neither beginning nor ending at the same position on the ordinate axis.

(3) By expressing the items of the respective series as percentages of the first or last amount. This method of treating different series, as shown in Figure 51, has the effect (*a*) of beginning or ending, as the case may be, the different curves *at the same positions on the ordinate axis*, and (*b*) of making the nature and the amount of deviation in the different series, as well as in the same series, directly comparable with each other, since the base amounts are treated as equal—100 per cent—in computing the percentages. It has the disadvantage that (*a*) relative rather than absolute amounts are plotted, (*b*) the curves may lie too close together, and (*c*) the first or the last item may not be suitable as a base.

FIGURE 51

Capital and Clearings of New York Clearing House Banks, 1902-1915

(Method of Scale Conversion)



Adjustments such as those described above are necessary only when the ordinate scale shows actual amounts and differences. When ratio changes are illustrated, they are unnecessary, because at any position on the ordinate axis equal ratios are indicated by equal spaces. A hundred per cent increase, whether representing the change from 2 to 4, 4000 to 8000, or 250,000 to 500,000, etc., always takes the same vertical space.

Charts designed to show ratio changes are discussed in the section immediately following.

If the time intervals are different in two series, and both are to be placed upon the same chart, an adjustment of the abscissa scale is necessary. In all cases, however, equal units on this axis should represent equal *periods* of time or *instants* equally distant apart. If, for example, one series is given by months, and another one by years, the same space cannot be allotted to both periods. If this were done, the time changes in *each* series but not those in *different* series would be comparable.

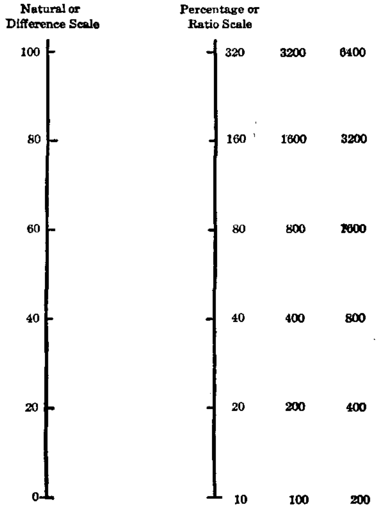
### b. Ratio Scales and "Ratio" Charts

Ordinate axes may show either actual or ratio changes in time series. If the former, equal spaces will indicate equal differences, positive or negative; if the latter, they will show equal rates of change. But spaces on an ascending scale indicating a given rate of increase do not show on a descending scale the same rate of decrease. That this is so may he seen from a simple example. A change from 100 to 200 represents an increase of 100 per cent, but a change from 200 to 100 is a decrease of 50 per cent. The reason for the difference is that, in the first case, the base is 100; in the latter, 200. That is, different bases are used in computing increases and decreases.

Comparable "difference" and "ratio" scales—arithmetic and geometric progressions—are shown in Figure 52.

FIGURE 52

A Natural or Difference Scale Contrasted with a Percentage or Ratio Scale

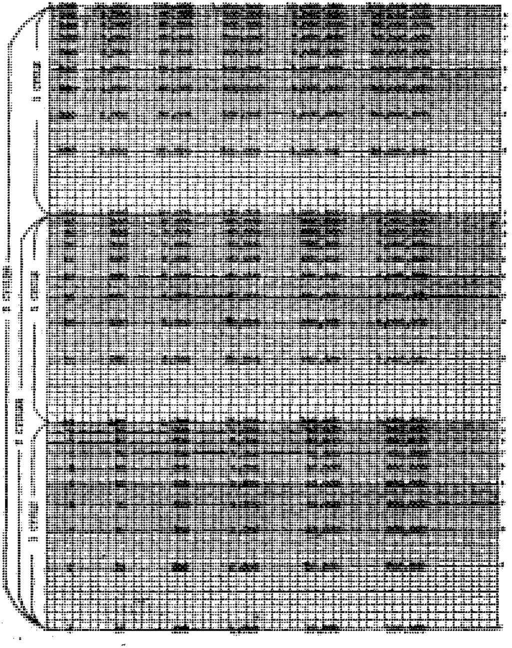



Rates of changes may be shown graphically in either of wo ways: (1) by plotting the logarithms of the amounts on difference scale, or (2) by plotting the amounts themselves n a logarithmic or ratio background. The latter alternative s simpler and preferable because (1) the meaning of logarithms of numbers is not generally understood, and (2) speially prepared paper is available upon which ratio changes

may be plotted, while log tables are not always accessible. Ratio paper is prepared in a variety of forms of which the following is an illustration.

FIGURE 53

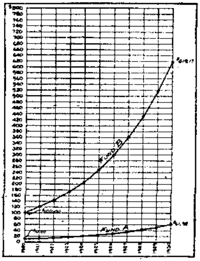
Illustration of How Different Scales May Be Placed on a Ratio Background



Alternative methods of showing the same facts (1) on a "difference," and (2) on a "ratio" [1] basis, are given in Figures 54-55.[2]

FIGURE 54 FIGURE 55

DIFFERENCE AND RATIO CHARTS SHOWING THE CHANGES IN FUNDS "A" AND "B"



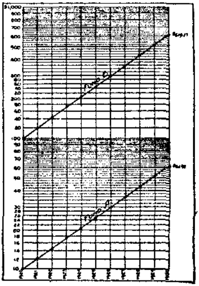


Figure 54 shows two series—A and B—plotted on ordinary arithmetic rulings—equal spaces representing equal amounts. From the figure it appears that the rate of increase in series "B" is more rapid than in series "A." This, however, is not the case as is shown in Figure 55, in which the series are drawn on a ratio background. Twenty per cent each year is added to the items in both series. The uniform rate of increase is properly brought out in the ratio chart, Figure 55.

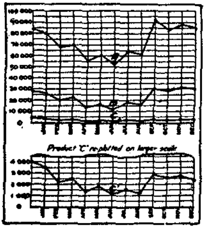
[1] Ratio paper in different sizes may be secured, among others, from The Education Exhibition Company, New York; Keuffel and Esser, Chicago and New York; Standard Graph Company, New York; Codex Book Company, New York.

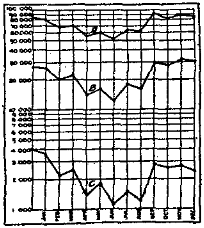
[2] The figures are reproduced with permission from Bivins, P. A., "The Ratio Chart and Its Applications," *The Engineering Magazine*, New York, July 1, 1921, p. 2 (Reprint).

Figures 56 and 57,[1] respectively, show the volume of sales of three products plotted on a difference and a ratio basis. On account of the limits of the scale, product "c" is plotted twice —the lower part of Figure 56 having a larger scale than the upper part. In Figure 57, the rates of movement of the respective products can be easily compared, two "cycles" of ratio ruling being used to show the movements. This chart illustrates the advantage of the ratio basis of showing amounts widely different in size. No complicated method of scale conversion is necessary, as is so often the case under such circumstances when a natural or difference scale is used.

FIGURE 56 FIGURE 57

DIFFERENCE AND RATIO CHARTS SHOWING THE CHANGES IN VOLUME OF SALES OF THREE PRODUCTS





The advantages of the "ratio" chart have been summarized by various writers,[2] but no more tersely than by Professor Irving Fisher. He says:

[1] *Ibid.*, p. 3.

[2] See Field, James A., "Some Advantages of the Logarithmic Scale in Statistical Diagrams," *Journal of Political Economy*, October, 1917, pp. 806-841. This article is reprinted in the author's *Readings and Problems in Statistical Methods*, Macmillan & Company, New York, 1920, pp. 282-305.

"The eye reads a ratio chart more rapidly than a difference chart or a table of figures. We may recapitulate what most easily catches the eye as follows:

"1. If we see a curve ascending, and nearly straight, we know that the statistical magnitude it represents is increasing at a nearly uniform rate.

"2. If the curve is descending, and nearly straight, the statistical magnitude is decreasing at a nearly uniform rate.

"3. If the curve bends upward, the rate of growth is increasing.

"4. If downward, decreasing.

"5. If the direction of the curve in one portion is the same as in some other portion it indicates the same percentage rate of change in both.

"6. If the curve is steeper in one portion than in another portion, it indicates a more rapid rate of change in the former than in the latter.

"7. If two curves on the same ratio chart run parallel they represent equal percentage rates of change.

"8. If one is steeper than another the first is changing at a faster percentage rate than the second.

"9. The imaginary straight line most nearly representing, to the eye, the general trend of the curve, is its 'growth axis,' and represents the average rate of increase (or decrease); and the deviations *of the curve from this growth axis are plainly evident without* recharting.

FIGURE 58

Domestic Orders for Freight Cars and Locomotives, Plotted on a Ratio Chart

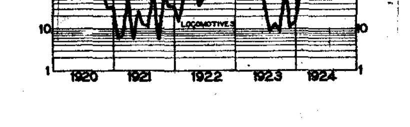


FIGURE 59

Rate of Turnover of Bank Deposits, Plotted on a Ratio Chart

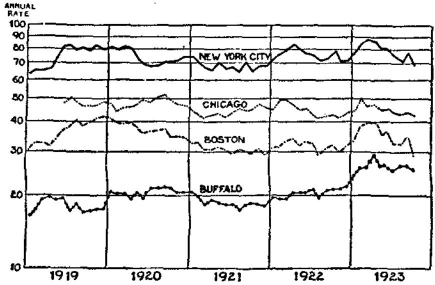
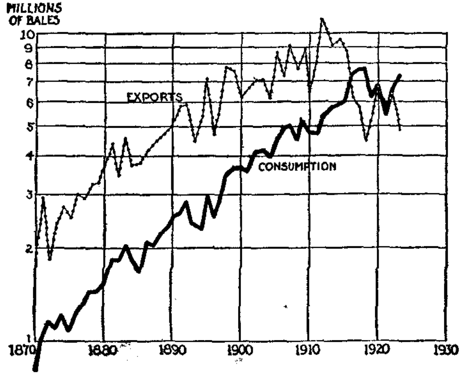



FIGURE 60

*Exports and Domestic Consumption of Cotton, Plotted on a Ratio Chart*



"10. The slope of the imaginary line between any two points on a curve indicates the average rate of change between the two."[1]

Figures 58, 59, and 60 are inserted to illustrate the different uses of ratio charts.

### (2) *Types of Lines Connecting Successive Ordinates*

Amounts or frequencies in historical series are generally *cumulated* through a period of time. This is the case, for instance, respecting exports, bank clearings, and industrial failures, reported by days, months, years. When they are plotted, the ordinates show what has been accomplished *during*, and not their characteristics *in*, such periods, deviations from which may be positive or negative. But the time intervals are arbitrary since time is continuous. The cumulations are functions of the periods selected in which to express the facts. Accordingly, while the amounts on the abscissa scale should be indicated as applying to the close of the periods in question, the respective ordinates should be connected by continuous smoothed lines. Such lines give a picture of the probable cumulations thought of as occurring through continuous time. Of course, if the periods are looked upon as discrete—which they are not—then a smoothed and continuous curve does not truly represent the facts. From this point of view, cumulation is begun anew at the beginning of each period and is completed at its end. But periods have neither beginnings nor endings except as arbitrarily conceived. To look upon them as discrete is absurd. Each ordinate is simply a conventional stopping place—it may be made earlier or later. If it is altered, then the cumulations are changed. Graphically, a continuous smoothed line shows the probable changes at all possible intervals comprehended in the entire period to which the data refer.

[1] **Fisher, Irving, "The 'Ratio' Chart for Plotting Statistics,"** ***Quarterly Publications of the American Statistical Association*****, June, 1917, pp. 597-598.**

Of course, too much liberty may be taken in drawing a smoothed line. The heights of the ordinates should be closely followed if the smoothed curve is taken to represent the probable cumulations of the case in question. If it is intended to represent an ideal cumulation, then the case is somewhat different. In the latter case, the question immediately arises: What is the "ideal" which is to be shown? Until it can be answered, smoothing should not be *too* "free hand."

On the other hand, certain historical series represent, not accumulations at the close of arbitrary periods, but characteristic facts, deviations being positive or negative, and coincident with the passage of time. Of such a nature are those relating to changes in temperature, barometric pressure, ratios of expenses to sales and of assets to liabilities, turnovers of bank deposits, etc. For such series, ordinates should be erected at the *middle* points of the time-units and be connected by smoothed lines. In reality, they are composed of a succession of continuous frequency series, because not only time but also the measurements are continuous. The units on both axes are arbitrary and artificial. Under such circumstances, smoothed curves give more than a direction of trend: they idealize both the units and the measurements.

When related series are plotted on the same chart, they should be designated by similar but distinguishable lines. On the other hand, lines which lie closely together or frequently cross each other should be drawn so as not to be confused. Since the use of lines of different color is generally prohibitive in cost, it is necessary to choose distinctive types of the same color where many curves are drawn upon one sheet. Lines should be broad enough to be readily followed, but not so broad as to sacrifice the accuracy of the ordinate unit.

### (3) *Purposes and Methods of Smoothing Historical or Time Series*

The methods used to smooth historical or time series depend upon the purposes to be accomplished thereby. The two

major purposes are (1) to secure a general notion of direction or trend, and (2) to analyze trends into their component parts as preliminaries to comparisons. Changes in time may be of four different types: (1) long-time or secular, (2) seasonal, (3) cyclical or periodic, and (4) "residual"—a term meant to cover all "other" types. Different methods of treating time series so as to isolate the first three classes of movements are discussed later in Chapter XIV.[1] The discussion at this point has to do with the first purpose.

If nothing more than a knowledge of general direction is desired, the free-hand method may suffice. If it is inadequate, the method of "moving averages" or "progressive means" may be used in series which are cyclical or periodic. This method involves (1) fixing approximately the length of the cycle, (2) totaling the frequencies or amounts for the first complete cycle, and taking the arithmetic average, (3) dropping off the first and adding a new item, totaling the amounts, and taking the arithmetic average, (4) continuing this process until the entire series is exhausted, (5) plotting the different averages at the middle points of each of the cycles, if they contain an even number, or half-way between the middle points if they contain an odd number of items.

This process, however, leaves the beginning and the end of the series unsmoothed. If the direction of the smoothed curve is fairly definite, however, the remaining parts of the series may be covered (1) by projecting the curve at both ends in keeping with its general inclination, or (2) by assuming that data similar to those at the respective ends of the series are repeated and by continuing to use moving averages.

This method, however, can be used with precision only when series are regularly cyclical or periodic. But how is this fact to be determined? Inspection often suffices to *suggest* a cycle but it does not *define* its exact length or its true periodicity. To secure a general direction of trend, however, it is not nec-

[1] *Infra*, pp. 441-457.

essary to have precise knowledge in either respect. If the approximate length of the cycle is used, moving averages will give, for general purposes, sufficiently accurate results. The more nearly it can be approximated, however, the better will be the results obtained.

If a period which corresponds to a half cycle, for instance, is used, the resulting curve, while it will smooth out the minor fluctuations of the incomplete periods, will not materially affect the longer changes. If a period somewhat shorter or longer is taken, the smoothed curve will partake of both the short- and long-time changes. In cases where periods are so dissimilar that a distorted curve is secured by using an average period, it is best not to employ the moving average method.

If historical series are to be correlated or minutely compared, then neither the free-hand nor the moving average method can be used. The trends must then be isolated. Different ways of doing this are discussed later.[1]

### 2. PLOTTING CUMULATIVE HISTORICAL OR TIME SERIES

Historical or time series relating to amounts or frequencies during a period of time may be cumulated. If, on the other hand, they have reference to characteristics of conditions at instants of time, they cannot be cumulated. Illustration will make the difference clear. If sales were available by months, the amounts at the successive intervals could be totaled so as to show the accumulation during any period of time. Sales in February could be added to those of January, and those of March to the combined total, etc., in the same way that successive frequencies are added in frequency series. On the other hand, temperature measurements at successive hourly intervals, ratios at different periods, etc., cannot be treated in this manner. To add or cumulate them is meaningless.

[1] See Chapter XIV, *passim*.

Successive ordinates in cumulated time series showing amounts should be connected by smooth continuous lines. Whatever the time unit used, it is arbitrary: continuity is suggested by an unbroken smooth line.

Ratio changes cannot be cumulated. To add and subtract successive ratios has no meaning. Moreover, a ratio chart is not suited to show cumulatively what has transpired. The scale showing increase has to be differently interpreted from that showing decrease.

## V. Conclusion

The discussion in this chapter has emphasized graphic as contrasted with diagrammatic presentation, attention being given primarily to (1) the distinction between discrete and continuous series and the manner in which they can be truly illustrated; (2) the processes of smoothing frequency series, and the meaning to be given to smoothed lines, (3) the methods of cumulating series and their graphic representation, (4) the use of difference and ratio scales in the graphic representation of time series, (5) scale conversion and rough methods of smoothing historical series, and (6) illustrations of types of graphic charts in current use.

Clear thinking about graphic representation and consistent use of devices for this purpose require that distinction be made between diagrams—pictorial illustrations—and lines and points fixed by a system of co-ordinates.

### References

American Telephone & Telegraph Company, New York,
"Graphical Method of Smoothing a Series of Frequency Curves," *Statistical Bulletin*, No. 2, April, 1921.
"Introduction to Graphic Methods, Part I," *Statistical Bulletin*, No. 3, November, 1921; Part II, No. 5, June, 1922.

Bivins, Percy A., "The Ratio Chart and Its Applications," *The Engineering Magazine*, The Engineering Magazine Company, New York, July, August, September, October, 1921.

BOWLEY, A. L., *Elements of Statistics*, King, London, 1911, Chapter VII, Sections I, II, III, IV, pp. 143-188.

BRINTON, W. C., "Graphic Methods for Presenting Facts," *Engineering Magazine*, New York, 1914, Chapters IX, X, pp. 149-163, 164-199, respectively.

ELDERTON, W. P., and ETHEL M., *Primer of Statistics*, Black, London, 1910, Chapter III, pp. 23-39.

FISHER, IRVING, "The 'Ratio' Chart for Plotting Statistics," in *Quarterly Publications of the American Statistical Association*, June, 1917, pp. 577-601.

HASKELL, ALLAN C., *Graphic Charts in Business*, Codex Book Company, New York, 1922, *passim*.

KARSTEN, K. G., *Charts and Graphs*, Prentice-Hall, New York, 1923, *passim*.

KELLEY, TRUMAN L., *Statistical Method*, Macmillan & Company, New York, 1923, Chapter II, pp. 9-48.

KING, W. I., *Elements of Statistical Method*, Macmillan & Company, New York, 1912, Chapter XI, pp. 97-120.

MARSHALL, ALFRED, "On the Graphic Method of Statistics" in the *Jubilee Volume, Journal of the Royal Statistical Society*, 1885.

MARSHALL, W. C., *Graphical Methods*, McGraw-Hill Book Company, New York, 1921, *passim*.

MILLS, FREDERICK C., *Statistical Methods Applied to Economics and Business*, Holt, New York, 1924, Chapter II, pp. 11-60.

THORNDIKE, E. L., *An Introduction to the Theory of Mental and Social Measurements*, Columbia University, New York, 1916, Chapter III, pp. 28-41.

YULE, G. U., *An Introduction to the Theory of Statistics*, Griffin, London, 1911, Chapter VI, pp. 75-105.

# CHAPTER IX

## AVERAGES AS TYPES

### I. Introduction

The discussion in the previous chapters, like Gaul, may be divided into three parts. Chapter I defines the subject matter of the book; Chapters II to V, inclusive, describe the manner in which statistics are assembled and collected from secondary and primary sources, respectively; while Chapters VI to VIII, inclusive, discuss the ways in which data are arranged in tables, and illustrated by diagrams and graphs. The discussion has to do with the processes of securing and arranging series of statistical aggregates rather than with the constants which may be used to describe them; it relates more to the manner in which they are built up than to the relations between the different parts; more to them in the gross, than in the net; more to them as details, than as summaries.

Statistical aggregates make up series of one type or another, descriptive of complex phenomena in point of time, space, or condition. The phenomena with which they deal are "affected to a marked extent by a multiplicity of causes"; they do not stand alone.[1] If they are to be adequately described by statistics, then the processes to which so much attention has been given in the foregoing chapters must be carried out with scrupulous care.

Statistical series, however, can rarely be adequately dealt

[1] "Life and the social process are not made up of bracketed situations of cause and effect, means and ends, stimulus and response. On the contrary, life is composed of related and interrelated situations * * * *Life is flow, process. The real search is not for action and reaction,* but interaction." Lindeman, Eduard C., *Social Discovery*, Republic Publishing Company, New York, 1924, p. 44.

with without using some kinds of summaries. Comparisons make them imperative. Expressions which are descriptive of the characteristics of data are required. Averages of various types serve this purpose or function.[1] The mind craves some sort of an average when dealing with series of statistical facts. Interest may be in the average price, the average student, average sale, average "business conditions" or what not, when dealing with phenomena of these types. Relations must be established, and for this purpose the details of series are too involved. They must be reduced to a single expression which stands for or reduces them to a unit basis.[2]

Averages are used loosely in everyday life. They often serve as a cloak for ignorance—people being willing to summarize their opinions in this way when they have no information concerning either the function of an average or the detail which it summarizes. They are used to give general impressions expressive of one's prejudices, general notions, sympathies or feelings of what ought to be the case in particular situations. "Short cuts" of this type are used in making broad generalizations about affairs for which often no average is available, and which cannot be summarized in this manner. Averages are the chief stock in trade of those who are loose minded, and prone to generalize. The expression, "on the average," is greatly overworked—so much so that it is hackneyed. Its free use suggests, if it does not always indicate,

[1] "An average * * * in general we may regard as one of a class of statistical constants * * * which concisely label a set of observations or measurements pertaining to a common family. It is designed to describe the family type more nearly than is possible by observing any chance member, and in value it should therefore come somewhere near the middle of the family group, so that if the individual members of the family chance to be equal each to each in respect to the organ or character observed it should have the same value as they have." Jones, D. Caradog, *A First Course in Statistics*, Bell, London, 1921, p. 23.

[2] In speaking of the arithmetic average, Keynes says, "But the utility of an average generally consists in our supposed right to substitute, in certain cases, this single measure for the varying measures of which it is a function." Keynes, J. M., *A Treatise on Probability*, Macmillan & Company, Ltd., London, 1921, p. 205.

the unscientific mind. To be scientific is to be able to identify similarities and differences and to be precise in one's generalizations about them. The willingness always to use averages is not in keeping with this requirement.

(Rarely, if ever, does an average[1] contain as much significance as do the detailed data which it summarizes.[2] It is used as a substitute for that which it replaces, but in this fact lies its chief limitation. The same average amount may be computed from different details, yet it may be these which are of chief interest. If averages alone are used, then the details, except in so far as they are reflected in such summaries, are ignored. As the formulation of a physical or a natural law depends upon observation and experiment, so the use of an average grows out of analysis of statistical detail. It presupposes (1) a purpose, (2) a knowledge of the peculiarities of the data to be averaged, (3) a clear conception of the properties of the appropriate average, and (4) a mastery of the whole subject to which the data relate so as to be sure that the average selected will have the proper significance/ 2

## II. Common Averages Defined

The averages with which we are concerned are those in common use. They are as follows: (1) the *arithmetic mean* or *average,* (2) the *median,* (3) the *mode,* and (4) the *geometric mean.* At this stage of the discussion, definitions of each kind will suffice. Their peculiar properties and uses will be discussed later.

The *arithmetic mean* or *average* is the amount secured by dividing the sum of the values of the items in a series by their number.

[1] Watkins speaks of averages as "representative numbers" and as containing "the gist, if not the substance, of statistics." Watkins, G. P., "Theory of Statistical Tabulation," *Quarterly Publications of the American Statistical Association,* December, 1915, p. 752.

[2] Venn, Dr. John, "On the Nature and Use of Averages," *Journal of the Royal Statistical Society* (London), Vol. LIV, 1891, pp. 429-448, at p. 433.

The *median* of a series is the value of that item—actual or estimated—when a series is arranged in order of magnitude, which divides the distribution into two equal parts.

The *mode* of the items in a series is the value of the one or ones which are most characteristic or common. It is the typical fact and always relates to a condition which is actually represented.

The *geometric mean* of the items in a series is the result secured by multiplying together the values of the various items and taking the *nth* root of their product.

These are all averages of the "first" order—that is, they have to do with the actual items in statistical series. In contrast to them, we shall later [1] consider averages of the "second" order—those which summarize not the actual items but the differences between them and some standard amount.)

## III. The Arithmetic Mean or Average

### 1. what it is

The arithmetic mean is the most familiar average in current use. Indeed, it is the only one customarily employed by the "man in the street." To him, *an* average is *the* average—the arithmetic mean about which he learned in his school days and about which, in its technical aspects, he has given little or no thought. Its use is a matter of daily routine in business. Why discuss it in a book on statistical methods! Common use and the assurance that it is fully understood do not, however, make a discussion of it unnecessary. It may appear that it is understood as to method of calculation, but not as to use and relation to other averages—matters about which little or nothing is commonly known.

According to definition, the arithmetic mean is the result secured by *adding* together the values of the items in a series and by *dividing* the total by the number of items. Thus, the

[1] Chapter X, *passim*.

arithmetic mean of 5 and 3 is secured by adding one 5 to one 3 and dividing by 2. The result, 4, is the average. The differences of the items from the average—plus and minus—are numerically equal, their algebraic sum being zero. In the illustration, 5 exceeds 4 by the same amount as 3 falls short of it. Accordingly, such a statistical constant is the center of gravity or point of balance of the items in a series. Moreover, it should be noted that in adding the quantities the influence of each upon the total is proportional to its size. On the other hand, in dividing the total by the number of its constituent parts, the items are treated as equal. Accordingly, the arithmetic mean is much influenced by the relative *size* of the items.)

Moreover, the *same* average amount may be secured from a variety of series. To illustrate: The arithmetic mean of 8, 9, 10, 11, 12, 13, and 14 is 11. So also is 11 the arithmetic mean of 8, 8, 8, 9, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 12, 13, 13, 13, 14, 14, 14; of 2 and 20; of 9, 9, 4, 22; of 3, 1, 1, 1, 1, 99, 1, 1, 1, 1, 11; and of many other combinations of items which might be selected. When an average is thus wholly independent of (1) the order of the items, (2) the number of items, and (3) their relative size, it has serious limitations for uses in which the nature of the distribution which is averaged is of interest. Moreover, this average may never be represented in a series. This is the case, for example, when 2 and 20; or 9, 9, 4, 22 are averaged. The result is always the center of gravity, but such a center may not represent an actual case. It is fictitious in this sense, although real in the sense that the product secured by multiplying it by the number of items gives the sum of the parts. Indeed, for the calculation of this average it is not necessary to know the size of the items provided the number and their total are given.

(If an average is to be taken as a substitute for detail, then the arithmetic mean, in spite of its simplicity and ease of calculation, has little to recommend it when series are

non-homogeneous. It is true that the average can be substituted for each item in a series, and the same total be secured, but substitution of this nature may not be wanted, the characteristic amounts being of interest. An arithmetic mean wage-rate, for instance, may tell the management of a plant the number of equal parts into which his wage bill is divided, but it does not show what the different employes actually receive. An arithmetic mean does not necessarily indicate the nature of the parts of which it is the center of gravity.

In the more precise measurements of the physical sciences its use is well established. "If we have $n$ observed values of an unknown, all equally good so far as we know, the most plausible value of the unknown (best value on the whole) is the arithmetic mean of the observed values."[1] Speaking further, the same writers say, "When the number of observed values is very great, the arithmetic mean is the true value."[2] This claim is based upon the principle that, in the absence of bias, large errors or deviations are less frequently encountered than are those which are small, the errors tending to be distributed about a true value according to the laws of probability or chance. That is, positive and negative deviations of the same size tend to occur with the same frequency.[3]

The fact that errors in measurements relating to economic and social phenomena are not subject solely to chance makes it impossible in such cases to use with assurance the arithmetic mean as the "true" average. Observations are not necessarily all "equally good." They are affected by the peculiarities of the units, personal bias, changing purposes, and varying motives. The ways in which these affect meas-

[1] Wright, T. W., and Hayford, J. F., *The Adjustment of Observations*, D. Van Nostrand, New York, 1906, p. 10.

[2] *Ibid.*, p. 11.

[3] Certain mathematical properties of the arithmetic mean are discussed by Yule, G. U., in *An Introduction to the Theory of Statistics*, Griffin, London, 1911, pp. 114 ff. and in Wright and Hayford, *op. cit.*, Chapter I.

urements of economic and social phenomena have already been discussed in earlier chapters.

### 2. HOW THE ARITHMETIC MEAN IS COMPUTED

The fact that the arithmetic mean of a series is its center of gravity is illustrated in Figure 61. The series of which the mean is to be calculated is given in Table 31.

TABLE 31

TABLE SHOWING WAGE-RATES AS BASES FOR THE COMPUTATION OF A SIMPLE ARITHMETIC MEAN RATE

| THE UNIT OR AMOUNT AVERAGED | THE NUMBER OF TIMES EACH UNIT IS ENCOUNTERED (The Weight) |
|---|---|
| $39.00 | 9 |
| 2.00 | 1 |
| 4.00 | 1 |
| 3.00 | 1 |
| 6.00 | 1 |
| 3.00 | 1 |
| 8.00 | 1 |
| 5.00 | 1 |
| 3.50 | 1 |
| 4.50 | 1 |

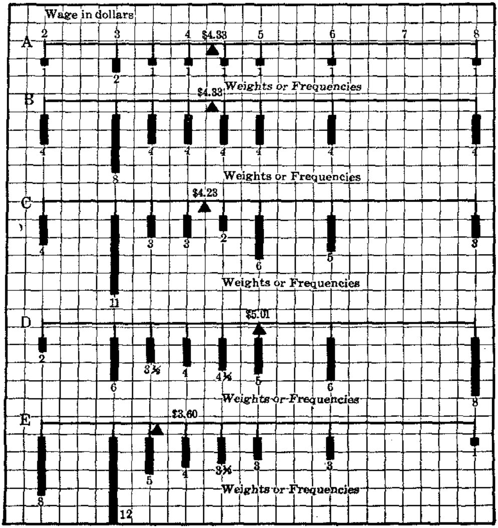
The sum of the values of the items, $39, divided by the number of items, 9, is $4.33. This is the arithmetic mean. If the different items are suspended as weights upon an imaginary rod, as in Figure 61, part A, the rod will balance at the scale unit $4.33. If, to the same units, frequencies (weights)[1] greater than unity but proportionally the same as in the first oase are assigned, the rod will balance at the same place. This adjustment is shown in part B of Figure 61. In this case, the frequencies (weights) have been multiplied through-

[1] See the discussion, *infra*, pp. 279-281, on the distinction between a simple and a weighted series.

out by 4: that is, *each* of them is made four times as heavy. If, however, the relations between the frequencies (weights) are changed, as they are in part C of the figure, then the average will change: that is, the center of gravity will be disturbed.

FIGURE 61

DIAGRAMS ILLUSTRATING THE NATURE OF THE ARITHMETIC MEAN WHEN ITEMS ARE DIFFERENTLY WEIGHTED




If the adjustment is made according to chance, the differences between the two results will be small. Frequencies (weights) of some sort are always present: the effect which

they have on the average is determined by their relative size and by their distribution. Taking the same units as above, and the chance frequencies (weights) given in Table 32, the average is reduced by only \$.10—that is, it is \$4.23—notwithstanding the fact that the difference between the extreme frequencies (weights) is 7, and that the frequency (weight) of one item is 4½ times as large as that of another.

TABLE 32

TABLE SHOWING WAGE-RATES WITH NUMBER OF PERSONS RECEIVING THEM AS A BASIS FOR COMPUTING AN ARITHMETIC MEAN RATE

| THE UNIT OR AMOUNT AVERAGED | THE NUMBER OF TIMES EACH UNIT IS ENCOUNTERED (The Weights) | PRODUCT OF THE WEIGHT TIMES THE UNIT |
|---|---|---|
| Total ............. | 37 | \$156.50 |
| 2.00 | 4 | 8.00 |
| 4.00 | 3 | 12.00 |
| 3.00 | 9 | 27.00 |
| 6.00 | 5 | 30.00 |
| 3.00 | 2 | 6.00 |
| 8.00 | 3 | 24.00 |
| 5.00 | 6 | 30.00 |
| 3.50 | 3 | 10.50 |
| 4.50 | 2 | 9.00 |

By arbitrarily adjusting the frequencies (weights) for each of the items, the average may be increased or decreased at will between the largest and smallest values. Column 1, Table 33, shows frequencies selected in such a manner that the values larger than the average (when all values are taken once) are given large frequencies (weights) and those smaller than the average small frequencies (weights), the importance varying directly with the size of the unit. In column 2, Table 33, the relative size of the frequencies (weights) is reversed. Diagrammatically, the effect of choosing such frequencies (weights) is shown in parts D and E, respectively, of Figure 61.

TABLE 33

TABLE SHOWING WAGE-RATES WITH NUMBER OF PERSONS RECEIVING THEM AS A BASIS FOR COMPUTING ARITHMETIC MEAN RATES

| THE UNIT OR AMOUNT AVERAGED | COL. 1 THE NUMBER OF TIMES EACH UNIT IS ENCOUNTERED (The Weights) | PRODUCTS OF UNITS AND WEIGHTS | COL. 2 THE NUMBER OF TIMES EACH UNIT IS ENCOUNTERED (The Weights) | PRODUCTS OF UNITS AND WEIGHTS |
|---|---|---|---|---|
| Total........ | 39 | $195.50 | 39.5 | $142.25 |
| $2.00 | 2 | 4.00 | 8 | 16.00 |
| 4.00 | 4 | 16.00 | 4 | 16.00 |
| 3.00 | 3 | 9.00 | 6 | 18.00 |
| 6.00 | 6 | 36.00 | 3 | 18.00 |
| 3.00 | 3 | 9.00 | 6 | 18.00 |
| 8.00 | 8 | 64.00 | 1 | 8.00 |
| 5.00 | 5 | 25.00 | 3 | 15.00 |
| 3.50 | 3½ | 12.25 | 5 | 17.50 |
| 4.50 | 4½ | 20.25 | 3½ | 15.75 |
| Average | | 5.01 | | 3.60 |

By thus arbitrarily selecting the frequencies (weights), the exact sizes being essentially within the limits of those assigned by chance, the resulting average is increased in the first case (column 1) over that secured by assigning equal frequencies by $.68, and over that gotten by assigning chance frequencies (weights) by $.78. In the second case, the average compared with that obtained by using equal frequencies (weights) is decreased by $.73, and when compared with that secured by using chance frequencies (weights) by $.63. The difference obtained by arbitrarily selecting the frequencies (weights) is $1.41 as compared with $.10 when equal and chance frequencies (weights) are used.

The arithmetic mean or average of a series of items is a function of the importance assigned to each one. It tends to be larger than the average of an equally weighted series when large items are heavily weighted, and smaller than it when

small items are heavily weighted. When frequencies (weights) are chosen at random, the resulting average is usually affected very little by their absolute size.

By taking the wage-rates above and assigning to them pure chance frequencies (weights)[1] (done by drawing by chance from a group of numbers marked with figures from 1 to 29, inclusive) the averages in four trials were found to be as follows: $4.43, $4.26, $4.29, and $4.04. These agree closely with the result secured when equal frequencies (weights) were used.

The commonly used method of computing arithmetic means is to total the values of the items and divide by the number of items. In some cases, however, particularly where there are many frequency groups and large items, it is easier to proceed in a different manner. In keeping with the principle that the sum of the deviations, signs considered, from the correct average equals zero, an average may be assumed as a starting point, the deviations calculated and corrected for error, and the correct result determined. This method of calculating an average for an ungrouped series of wage-rates is illustrated in Table 34. The trial average, $5, is assumed. The sum of the minus deviations = —$10; the sum of the plus deviations is $4; the algebraic sum is —$6. The trial average is, therefore, not the correct average. If it were, the algebraic

[1] The following are chance frequencies (weights) used in this experiment:

| Units | 1st Trial | 2d Trial | 3d Trial | 4th Trial |
|---|---|---|---|---|
| $2.00 | 25 | 22 | 13 | 23 |
| 4.00 | 22 | 24 | 21 | 14 |
| 3.00 | 17 | 11 | 23 | 6 |
| 6.00 | 23 | 26 | 24 | 28 |
| 3.00 | 1 | 27 | 14 | 15 |
| 8.00 | 15 | 16 | 10 | 1 |
| 5.00 | 27 | 16 | 20 | 10 |
| 3.50 | 12 | 25 | 19 | 2 |
| 4.50 | 21 | 23 | 24 | 3 |

(The student is advised to try others.)

sum would be zero. Since the net error is —\$6, the amount must be divided by 9, the number of instances, and the product algebraically added to \$5. The operation is as follows: $\frac{-\$6}{9} = -\$.67$. $\$5.00 + (-\$.67) = \$4.33$, which is the correct average.

TABLE 34

TABLE GIVING DATA FOR COMPUTING THE ARITHMETIC MEAN BY THE "SHORT-CUT" METHOD

| UNITS OR AMOUNTS | FREQUENCIES | DEVIATIONS | | |
|---|---|---|---|---|
| | | — | + | NET DEVIATIONS |
| Total | 9 | \$10.00 | \$4.00 | — \$6.00 |
| \$2.00 | 1 | 3.00 | | |
| 4.00 | 1 | 1.00 | | |
| 3.00 | 1 | 2.00 | | |
| 6.00 | 1 | | 1.00 | |
| 3.00 | 1 | 2.00 | | |
| 8.00 | 1 | | 3.00 | |
| 5.00 | 1 | | | |
| 3.50 | 1 | 1.50 | | |
| 4.50 | 1 | .50 | | |

The same method is followed in series in which the frequencies are greater than unity. The only additional step involved is to multiply the deviations by their respective frequencies. This is necessary because the deviations appear as many times as the items are encountered.

This would be apparent at once, if, instead of indicating the number of times each item appears, the alternative plan were followed of repeating the item itself. In Table 35, the process of calculating a mean in this manner is carried out in detail.

The total net deviation from the assumed average, \$5, is —\$93.50. That is, \$5 is greater than the true average. Accordingly, the total net error must be distributed over the 163

TABLE 35

TABLE GIVING DATA FOR COMPUTING THE ARITHMETIC MEAN BY THE "SHORT-CUT" METHOD

| UNITS OR AMOUNTS | FRE-QUENCIES | DEVIATIONS | | DEVIATIONS TIMES THE FREQUENCIES | | TOTAL NET DEVIATIONS |
|---|---|---|---|---|---|---|
| | | − | + | − | + | |
| Total | 163 | | | $161.50 | $68.00 | — $93.50 |
| $2.00 | 25 | $3.00 | | 75.00 | | |
| 4.00 | 22 | 1.00 | | 22.00 | | |
| 3.00 | 17 | 2.00 | | 34.00 | | |
| 6.00 | 23 | | $1.00 | | 23.00 | |
| 3.00 | 1 | 2.00 | | 2.00 | | |
| 8.00 | 15 | | 3.00 | | 45.00 | |
| 5.00 | 27 | | | | | |
| 3.50 | 12 | 1.50 | | 18.00 | | |
| 4.50 | 21 | .50 | | 10.50 | | |

items, and the result be algebraically added to **$5.** The computations involved are as follows: **—$93.50 ÷ 163 = —$.57. $5.00 + (—$.57) = $4.43,** which is the arithmetic mean.

When arithmetic means are to be computed for series which are grouped, some assumption must be made as to the size of the items in the respective groups. The conventional method is to assume that the frequencies in each group are distributed uniformly throughout its range, or, what amounts to the same thing, that they are concentrated at the center. How correct this is, for discrete and continuous series, has already been considered. In the absence of exact values, however, since precise amounts must be used, the conventional method may be followed.

The ordinary way of computing the arithmetic mean for a grouped series is shown in Table 36, the respective frequencies being multiplied by the central values of the groups.

TABLE 36

TABLE GIVING DATA FOR COMPUTING AN ARITHMETIC MEAN FROM FREQUENCY GROUPS

| UNITS OR AMOUNTS | FREQUENCIES | PRODUCTS OF FREQUENCIES AND THE UNITS (Middle Terms) |
|---|---|---|
| Total .................. | 434 | $3,923.00 |
| $5.00 to $5.99 | 15 | 82.50 |
| 6.00 to 6.99 | 40 | 260.00 |
| 7.00 to 7.99 | 66 | 495.00 |
| 8.00 to 8.99 | 91 | 773.50 |
| 9.00 to 9.99 | 113 | 1,073.50 |
| 10.00 to 10.99 | 49 | 514.50 |
| 11.00 to 11.99 | 30 | 345.00 |
| 12.00 to 12.99 | 27 | 337.50 |
| 13.00 to 13.99 | 2 | 27.00 |
| 14.00 to 14.99 | 1 | 14.50 |

$3,923 ÷ 434 = $9.04 = arithmetic mean or average.

TABLE 37

TABLE GIVING DATA FOR COMPUTING AN ARITHMETIC MEAN BY THE "SHORT-CUT" METHOD FOR FREQUENCY GROUPS FROM AN ASSUMED AVERAGE

| UNITS OR AMOUNTS | FREQUENCIES | DEVIATIONS FROM THE ASSUMED AVERAGE, $9.50 | | PRODUCTS OF DEVIATIONS AND FREQUENCIES | | NET DEVIATIONS |
|---|---|---|---|---|---|---|
| | | − | + | − | + | |
| Total ........ | 434 | | | $403.00 | $203.00 | — $200.00 |
| $5.00 to $5.99 | 15 | $4.00 | | 60.00 | | |
| 6.00 to 6.99 | 40 | 3.00 | | 120.00 | | |
| 7.00 to 7.99 | 66 | 2.00 | | 132.00 | | |
| 8.00 to 8.99 | 91 | 1.00 | | 91.00 | | |
| 9.00 to 9.99 | 113 | | | | | |
| 10.00 to 10.99 | 49 | | $1.00 | | 49.00 | |
| 11.00 to 11.99 | 30 | | 2.00 | | 60.00 | |
| 12.00 to 12.99 | 27 | | 3.00 | | 81.00 | |
| 13.00 to 13.99 | 2 | | 4.00 | | 8.00 | |
| 14.00 to 14.99 | 1 | | 5.00 | | 5.00 | |

If the method of computing the deviations from an *assumed* average is used, the steps are the same as those used when data are not arranged in groups, except that it is necessary, as in the case immediately above, to assume a uniform distribution throughout each group. The method is shown in Table 37, the trial average being $9.50, i.e., the item half-way through the group, $9.00 to $9.99.

— $200 ÷ 434 = — $.46. That is, the net average deviation does not equal zero, but —$.46. Therefore, in order to determine the true average (from which the sum of the deviations equals zero) it is necessary to add —$.46 to the assumed average, $9.50, thus giving $9.04 as the correct average.

The plus and minus deviations, calculated in the same manner *but from the actual average, $9.04*, are given in Table 38.

TABLE 38

TABLE SHOWING THE EFFECT OF COMPUTING THE ARITHMETIC MEAN FROM THE TRUE AVERAGE FOR DATA IN FREQUENCY GROUPS

| Units or Amounts | Frequencies | Deviations from the True Average, $9.04 | | Products of Deviations and Frequencies | | Net Deviations |
|---|---|---|---|---|---|---|
| | | — | + | — | + | |
| Total ........ | 434 | | | $305.48 | $305.12 | —$.36 * |
| $5.00 to $5.99 | 15 | $3.54 | | 53.10 | | |
| 6.00 to 6.99 | 40 | 2.54 | | 101.60 | | |
| 7.00 to 7.99 | 66 | 1.54 | | 101.64 | | |
| 8.00 to 8.99 | 91 | .54 | | 49.14 | | |
| 9.00 to 9.99 | 113 | | $ .46 | | 51.98 | |
| 10.00 to 10.99 | 49 | | 1.46 | | 71.54 | |
| 11.00 to 11.99 | 30 | | 2.46 | | 73.80 | |
| 12.00 to 12.99 | 27 | | 3.46 | | 93.42 | |
| 13.00 to 13.99 | 2 | | 4.46 | | 8.92 | |
| 14.00 to 14.99 | 1 | | 5.46 | | 5.46 | |

* This negligible difference is due to the fact of taking the average at $9.04. The exact average is $9.039 +.

When frequency groups are all of equal size, it is often a saving of time to compute the deviations from an assumed average in terms of the "steps" which successive groups are above or below the group containing the assumed average, and later to convert the net "step-deviations" back into real deviations by multiplying by 1, in case the step is unity, 2 in case it is two, by ½ in case it is one half, etc. Using the distribution in Table 38, but assuming a different average, the arithmetic mean is computed by the "step" method in Table 39.

TABLE 39

TABLE GIVING DATA FOR COMPUTING THE ARITHMETIC MEAN BY THE "STEP-DEVIATION" METHOD FOR FREQUENCY GROUPS FROM AN ASSUMED AVERAGE

| UNITS OR AMOUNTS | FREQUENCIES | "STEP-DEVIATIONS" FROM THE ASSUMED AVERAGE, $12.50 | | PRODUCTS OF "STEPS" AND FREQUENCIES | | NET "STEP-DEVIATIONS" |
|---|---|---|---|---|---|---|
| | | − | + | − | + | |
| Total ....... | 434 | | | 1506 | 4 | — 1502 |
| $ 5.00 to $ 5.99 | 15 | 7 | | 105 | | |
| 6.00 to 6.99 | 40 | 6 | | 240 | | |
| 7.00 to 7.99 | 66 | 5 | | 330 | | |
| 8.00 to 8.99 | 91 | 4 | | 364 | | |
| 9.00 to 9.99 | 113 | 3 | | 339 | | |
| 10.00 to 10.99 | 49 | 2 | | 98 | | |
| 11.00 to 11.99 | 30 | 1 | | 30 | | |
| 12.00 to 12.99 | 27 | | | | | |
| 13.00 to 13.99 | 2 | | 1 | | 2 | |
| 14.00 to 14.99 | 1 | | 2 | | 2 | |

— 1502 ÷ 434 = −3.46. −3.46 × $1.00 (the size of the group) = —$3.46. $12.50 (the assumed average) + (—$3.46) = $9.04 = the true average.

Where groups are not uniform in size, this method cannot be employed without considerable difficulty. When they are uniform, however, multiplying is simplified by computing the deviations in round numbers. The deviations, however, are

TABLE 40

TABLE GIVING DATA FOR COMPUTING THE ARITHMETIC MEAN BY THE "STEP-DEVIATION" METHOD FROM AN ASSUMED AVERAGE WHEN THE GROUPS ARE OF UNEQUAL SIZE *

| Groups | | | Frequencies | "Step-Deviations" | | Products of "Steps" and Frequencies | | Net "Step-Deviations" |
|---|---|---|---|---|---|---|---|---|
| Size | Width | Center | | − | + | − | + | |
| Total | | | 30,454 | | | | | |
| Total | | | 24,885 | | | 13,976 | 15,242 | +1266 ‡ |
| † Less than 6¢ | 2 | 5 | 99 | 4 | | 396 | | |
| 6¢– 8¢ | 2 | 7 | 661 | 3 | | 1,983 | | |
| 8¢–10¢ | 2 | 9 | 2,722 | 2 | | 5,444 | | |
| 10¢–12¢ | 2 | 11 | 6,153 | 1 | | 6,153 | | |
| (1) 12¢–14¢ | 2 | 13 | 6,007 | | | | | |
| 14¢–16¢ | 2 | 15 | 4,926 | | 1 | | 4,926 | |
| 16¢–18¢ | 2 | 17 | 2,635 | | 2 | | 5,270 | |
| 18¢–20¢ | 2 | 19 | 1,682 | | 3 | | 5,046 | |
| Total | | | 5,076 | | | 2,604 | 468 | —2136 § |
| 20¢–25¢ | 5 | 22.5 | 2,604 | 1 | | 2,604 | | |
| (2) 25¢–30¢ | 5 | 27.5 | 2,004 | | | | | |
| 30¢–35¢ | 5 | 32.5 | 468 | | 1 | | 468 | |
| Total | | | 291 | | | | | |
| (3) 35¢–45¢ | 10 | 40 | 291 | | | | | ‖ |
| Total | | | 202 | | | 109 | 33 | — 76 ¶ |
| 45¢–60¢ | 15 | 52.5 | 109 | 1 | | 109 | | |
| (4) 60¢–75¢ | 15 | 67.5 | 60 | | | | | |
| † 75¢ and over | 15 | 82.5 | 33 | | 1 | | 33 | |

* Data taken from Report of the Tariff Board on Schedule "K," Vol. IV., Part 5. *House Doc. 342, 62d Congress, 2d Session*, p. 997.

† Width of group assumed to be the same as that of the class to which it belongs.

‡ + 1266 ÷ 24,885 = .0509. .0509 × 2¢ (the width of the group) = \$.001018. \$.13 + \$.001018 = \$.1310 (average of the first group).

in "steps," and they must be converted into the units of the series by multiplying them by the appropriate factor. The group in this case is **$1.00**, hence the factor is **$1.00**.

Table 40 illustrates the method to be used when groups are of unequal size. In such cases it is generally simpler to proceed in the regular manner by multiplying through in the first instance.

### 3. SOME "DO'S AND DON'TS" IN THE USE OF AVERAGES

#### *(1) Do Not Average Averages Unless They Are Properly Weighted*

Example "A"

It is desired to secure the arithmetic average of the following series separately and combined:

Series 1: $3, $4, $4, $5; Series 2: $2, $6, $7.
Computation, Series 1: $3 + $4 + $4 + $5 = $16. $16 ÷ 4 = $4.
Computation, Series 2: $2 + $6 + $7 = $15. $15 ÷ 3 = $5.
Computation, Combined Series, *Correct:* $3 + $4 + $4 + $5 + $2 + $6 + $7 = $31. $31 ÷ 7 = $4.43.
Computation, Combined Series, *Incorrect:* $4 + $5 = $9. $9 ÷ 2 = $4.50.

(Notes to Table 40, continued)

§ — 2136 ÷ 5076 = — .421. — .421 × 5¢ (the width of the group) = — $.02105. $.275 + (— $.02105) = $.254 (average of the second group).
|| $.40 is the average of the third group.
¶ — 76 ÷ 202 = — .376. — .376 × 15¢ (the width of the fourth group) = — $.05640. $.675 + (— $.05640) = $.6186 (average of the fourth group).

| GROUPS | AVERAGES | WEIGHTS | PRODUCTS OF WEIGHTS AND AVERAGES |
|---|---|---|---|
| Total | $.1573 | 30,454 | $4790.5962 |
| (1) | .1310 | 24,885 | 3259.9350 |
| (2) | .2540 | 5,076 | 1289.3040 |
| (3) | .4000 | 291 | 116.4000 |
| (4) | .6186 | 202 | 124.9572 |

Example "B" [1]

It is desired to compute the average percentage relation of ren to sales for the experience shown in the following table:

| Net Sales (in 000's) | Total Sales | Total Rent | Per Cent Rent to Sales |
|---|---|---|---|
| Under $40 | $8,471,952 | $255,845 | 3.02 |
| $40 to 80 | 20,719,729 | 545,733 | 2.63 |
| 80 to 180 | 26,232,605 | 729,026 | 2.78 |
| 180 and over | 30,555,976 | 737,008 | 2.41 |
| Total ......... | $85,980,262 | $2,267,612 | 2.64 |

*Correct* method: $2,267,612 ÷ $85,980,262 = 2.64 per cent.

*Incorrect* method: $\frac{3.02 + 2.63 + 2.78 + 2.41}{4} = 2.71$ per cent.

### (2) *Do Not Confuse Simple and Weighted Arithmetic Averages*

An arithmetic average computed from series in which th frequencies are greater than unity is not *necessarily* weighted

a. Computation of *Simple* Arithmetic Averages for Serie (1) in Which the Frequencies Are Unity in Each Case, an (2) in Which They Are Greater than Unity

"A"

| Wage-Rates | Number | Product: Number Times Rate |
|---|---|---|
| $5 | 1 | $5 |
| 6 | 1 | 6 |
| 7 | 1 | 7 |
| 8 | 1 | 8 |
| 9 | 1 | 9 |
| Total.. | 5 | $35 |

Average = $35 ÷ 5 = $7.

"B"

| Wage-Rates | Number | Product: Number Tim Rate |
|---|---|---|
| $5 | 2 | $10 |
| 6 | 3 | 18 |
| 7 | 1 | 7 |
| 8 | 2 | 16 |
| 9 | 2 | 18 |
| Total.... | 10 | $69 |

$69 ÷ 10 = $6.90.

[1] See Secrist, Horace, "A Statistical Paradox" in *Journal of the Ame ican Statistical Association*, June, 1923, pp. 776-780.

The two series are in reality the same since Type "B" may be written in the form of Type "A" as follows:

| Wage-Rates | Number |
|---|---|
| $5 | 1 |
| 5 | 1 |
| 6 | 1 |
| 6 | 1 |
| 6 | 1 |
| 7 | 1 |
| 8 | 1 |
| 8 | 1 |
| 9 | 1 |
| 9 | 1 |
| Total.......$69 | 10 |

$\$69 \div 10 = \$6.90.$

b. Computation of *Weighted* Arithmetic Averages

A *weighted* arithmetic average is one secured by applying to the items weights determined by some evidence of importance other than that associated with the items themselves.[1]

*Example 1*

| Per Cent of Total Acreage | Relative Condition of Crop | Product of Per Cent Acreage and Relative Condition |
|---|---|---|
| 7/10 | good = 2 | 14/10 |
| 2/10 | fair = 3 | 6/10 |
| 1/10 | poor = 5 | 5/10 |
| Total............ | ............ | 25/10 |

Average condition $= 25 \div 10 = 2.5.$

[1] "The multiplying of a score by the number of cases having it has at times been called weighting, but in this text the term will be used to mean the multiplying of scores by amounts determined not at all, or not solely, by the population, but from other evidences of importance." Kelley, T. L., *Statistical Method*, Macmillan & Company, New York, 1923, p. 68.

*Example 2*

| Types of Employes | Number on Payroll | Relative Productivity | Productivity Index times Number |
|---|---|---|---|
| Men .......... | 5 | 1 | 5 |
| Women ........ | 4 | ¾ | 3 |
| Youths ........ | 3 | ½ | 1½ |

Total men-equivalents = 9½.

*Example 3*

| Family Budget Item | Relative Importance in Family Budget "The Weights" | Per Cent Increase | |
|---|---|---|---|
| | | July, 1914 to November, 1920 | Multiplied by Weights |
| Food .................. | 43.1% | 93 | 4008.3 |
| Shelter ................ | 17.7% | 66 | 1168.2 |
| Clothing ............... | 13.2% | 128 | 1689.6 |
| Fuel & Lighting......... | 5.6% | 100 | 560.0 |
| Sundries ............... | 20.4% | 92 | 1876.8 |
| Total ............... | 100.0% | ... | 9302.9 |

Average = 9302.9 ÷ 100 = 93.03 per cent.

### (3) *Distinguish Between Including and Not Including "Zero" Cases in an Average* [1]

| | ZERO CASES INCLUDED | ZERO CASES NOT INCLUDED |
|---|---|---|
| Average tariff duty [2] = | Amount of duty collected / Value of imports | Amount of duty collected / Value of imports paying duty |
| Average daily wage = | Total wages paid per year / Number of days in a year | Total wages paid / Number of full days worked for which wages were paid |
| Average amount of taxes paid = | Total taxes / Number of people | Total taxes / Number of tax payers |
| Average consumption of liquor = | Liquor consumed / Total population | Liquor consumed / Total number of consumers |
| Average number of accidents per day [3] = | Number of accidents / Number of days | Number of accidents / Number of days on which accidents occurred |

[1] See *supra*, pp. 80-81, 89, for a discussion of an analogous problem relative to statistical ratios or coefficients.

[2] See Secrist, Horace, *Readings and Problems in Statistical Methods*, Macmillan & Company, New York, 1920, pp. 334-341.

[3] *Ibid.*, pp. 164-184.

### 4. SUMMARY

In summarizing the discussion of the arithmetic mean, attention should be called to the fact that it is (1) easily understood, (2) readily calculated, (3) in everyday use, and (4) affected by all of the items in a series. Indeed, when nothing more is wanted, as a summarizing expression, than the total divided by the sum of the parts, it thoroughly meets the need. But in statistical analysis of economic problems requirements generally run far beyond this. Details, as well as averages, or at least averages other than the arithmetic mean, are required. It is to a discussion of these to which attention is now turned.

## IV. The Median

### 1. WHAT THE MEDIAN IS

The median of a series has been defined as the value of that item—actual or estimated—when a series is arranged in order of magnitude which divides the number of frequencies into two equal parts. It is in the nature of an average, but in fact is a "partition expression," being the value of the middle item when series are arranged in order of size. It may or may not be representative of the different values. As to whether it is or is not depends upon the nature of the distribution involved. Moreover, unlike the arithmetic mean, the sum of the amounts is not secured—except in "normal" distributions where the arithmetic mean and the median are the same—by multiplying the median by the number of items. The amounts are not added and averaged; they are arrayed. Again, in calculating it, each item, whether large or small, is assigned the same importance, all frequencies being treated alike. The exact size of all of the items except the median one may be unknown, and yet it can be determined, because the only requirement for its calculation is that the items be arrayed in order of magnitude and the center one chosen. Moreover, like

the arithmetic mean, the median may be a value not found in a series—it may be an estimated rather than an actual amount.

### 2. HOW THE MEDIAN IS DETERMINED

Since the median is the value of the middle item in a series, it is calculated by using the following formulæ: if the number of measurements, $n$, is odd, use $\frac{n+1}{2}$; if it is even, the median value lies between $\frac{n}{2}$ and $\left(\frac{n}{2}+1\right)$. Since, however, "the value of a measure is the value of its mid-point, this (the value of the measure at $\frac{n+1}{2}$) is equivalent to saying that the median is the limit of the range covered by $\frac{n}{2}$ measures counted either down from the top or up from the bottom."[1]

The manner in which the median is computed in an ungrouped series made up of an *odd* number of items is shown

TABLE 41

TABLE GIVING DATA FOR COMPUTING THE MEDIAN

| UNIT | FREQUENCIES |
|---|---|
| Total | 9 |
| $2.00 | 1 |
| 3.00 | 1 |
| 3.00 | 1 |
| 3.50 | 1 |
| 4.00 | 1 |
| 4.50 | 1 |
| 5.00 | 1 |
| 6.00 | 1 |
| 8.00 | 1 |

[1] Kelley, T. L., *Statistical Method*, Macmillan & Company, New York, 1923, pp. 55-56.

in Table 41. By using the data in Table 31, p. 267, but rearranging the units in an ascending order—an unnecessary step in computing the arithmetic mean—the series is shown in Table 41.

Applying the formula, $\frac{n+1}{2}$, when $n = 9$, we get $\frac{9+1}{2} = 5$, i.e., the fifth item divides the series into two equal parts. Counting down from the smallest item, or up from the largest one—a matter of indifference—$4.00 is found to be the median. It should be noticed that the total frequencies, rather than the range of the size of the items, are divided in half. In the illustration, $4.00 is only $2.00 away from the first item, but $4.CO away from the last. Moreover, in determining the median in this case, $2.00 is of as much importance as is $8.00. It is quite different, of course, respecting the arithmetic mean. Moreover, while retaining the frequencies as above, every item in the series except the middle one may be changed—the only limitation being that the order must remain ascending—and the median remain the same. Various adjustments of this type are given in Table 42.

The median in every case is the fifth item—$4.00. It is not affected at all by changing the *size* of the items above or below the fifth one so long as the *number* of items remains the same and the series is ascending. Indeed, it is not affected by the addition of other items provided as many less than the median as well as more than it are added. On the other hand, the arithmetic mean is determined by both the number and size of the items. The quantity $10,000 in column 6 has 5000 times as much influence as has the quantity $2.00 in determining the arithmetic mean. But they have equal influence in fixing the median since each one is represented once. The median, therefore, thought of as an average to be *substituted* for the different items in a series, may be used only when (1) the differences between the consecutive items are small, or (2) the series is of the normal law of error type, the items at or

TABLE 42

TABLE GIVING DATA SHOWING THE EFFECT OF CHANGES OF DISTRIBUTION ON THE MEDIAN AND THE ARITHMETIC MEAN

| FREQUENCIES | UNITS AND ILLUSTRATIONS | | | | | |
|---|---|---|---|---|---|---|
| Total 9 | 1st | 2d | 3d | 4th | 5th | 6th |
| 1 | $2.00 | $1.00 | $3.99 | $4.00 | $ .25 | $2.00 |
| 1 | 3.00 | 1.00 | 3.99 | 4.00 | .50 | 3.00 |
| 1 | 3.00 | 1.00 | 3.99 | 4.00 | .75 | 3.00 |
| 1 | 3.50 | 1.00 | 3.99 | 4.00 | 1.00 | 3.50 |
| 1 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 |
| 1 | 4.50 | 4.00 | 4.01 | 4.00 | 4.00 | 4.50 |
| 1 | 5.00 | 4.00 | 4.01 | 4.00 | 4.00 | 5.00 |
| 1 | 6.00 | 4.00 | 4.01 | 4.00 | 4.00 | 6.00 |
| 1 | 8.00 | 4.00 | 4.01 | 4.00 | 4.00 | 10,000.00 |
| Median | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 |
| Arith. Mean | 4.33 | 2.67 | 4.00 | 4.00 | 2.50 | 1,114.45 |

near the median being the most common. In the latter case, the median is the same as the arithmetic mean, deviations in excess and in defect of it tending to be distributed about a true value according to the law of chance. Under such conditions, it is as much the "true" average, in the mathematical sense, as is the arithmetic mean. But the two averages are rarely equal for the simple but sufficient reason that normal distributions are seldom, if ever, found.

When the number of items, $n$, in a series is even, the median lies between the $\frac{n}{2}th$ and $\left(\frac{n}{2} + 1th\right)$ items. If a series is discrete no actual case appears at such a position. If a median *amount* is selected it is purely arbitrary. If a series is continuous, each measure is an approximation to the true measure, and, theoretically, items appear between these limits. The conventional practice in both cases is to take an amount halfway between the middle items. The justification of doing this,

however, is different in the two types of series. For a series which is discrete, the median under such circumstances is fictitious; for one which is continuous, it is theoretically although not actually present in the series.

The calculation of the median of a series containing an even number of items may be illustrated by adding one item to each of the series in Table 42. For instance, if an item of $200 is added to the series in *Illustration* 1, the median becomes $3.75. That is, $n$ is now 10. The two formulæ giving the position of the median will then read as follows: $\frac{n}{2} = 5$; $\left(\frac{n}{2} + 1\right) = 6$. The median, therefore, lies between the 5th and the 6th item, that is, between $3.50 and $4.00. It is fixed conventionally at $3.75. If $8.00 is added to the same series, the median as located by these formulæ falls between $4.00 and $4.50. It may be arbitrarily given the value of $4.25. Moreover, if to the series in *Illustration* 2, $600, $10,000, $12,000, $13,000, and $14,000 are added, the median is still $4.00. In this case, however, the size of the median is the same as that of the adjacent items because they are identical.

When data are arranged in frequency groups, the problem of determining the median is the same as it is when they are not grouped, except that it is necessary arbitrarily to distribute the frequencies within the groups in order to interpolate for the exact median. What is wanted is not only the *median group,* but the *median item* in the group which divides a series in half. To express the units in groups rather than individually makes it necessary to approximate the value of each of them. For discrete series classified in narrow groups, and for all continuous series, the assumption of a uniform distribution is sufficiently accurate for most purposes. Any error arising from this assumption will be negligible.[1]

[1] This is more particularly true since at the median position the frequencies are generally numerous. This is always the case in distributions of the normal type and in those which approach it.

A grouped frequency series is shown in Table 43. In this case, $n$ is 434: that is, it is an even number. On the assumption that the items through the groups are uniformly dispersed, and that it is admissible to compute the exact median, the process is as follows: $\frac{n}{2} = 217$; $\left(\frac{n}{2} + 1\right) = 218$. The median, therefore, lies in the *group* containing the 217½th item. The value of this item is the median.

TABLE 43

TABLE GIVING FREQUENCY DATA FOR THE COMPUTATION OF THE MEDIAN

| UNITS OR AMOUNTS | FREQUENCIES |
|---|---|
| Total | 434 |
| $ 5.00 to $ 5.99 | 15 |
| 6.00 to 6.99 | 40 |
| 7.00 to 7.99 | 66 |
| 8.00 to 8.99 | 91 |
| 9.00 to 9.99 | 113 |
| 10.00 to 10.99 | 49 |
| 11.00 to 11.99 | 30 |
| 12.00 to 12.99 | 27 |
| 13.00 to 13.99 | 2 |
| 14.00 to 14.99 | 1 |

By counting down from the smallest item, the group $9.00 to $9.99 is found to contain all the items between 212 and 325. The 217½th man's wage-rate is, therefore, located within this group. On the assumption that the 113 men whose wage-rates fall within the group $9.00 to $9.99, inclusive, are uniformly distributed in the order of the size of their rates, the wage-rate which is half-way between that received by the 217th and the 218th man is $\frac{5\frac{1}{2}}{113} \times \$1.00$, or $.05 greater than $9.00, i.e., than

the amount received by the first man in this group.[1] This gives a median wage-rate of $9.05 which corresponds very closely to the arithmetic mean, $9.04, as computed for the same data—Table 36.

Since this example has to do with wage-rates—a discrete series—the median might with sufficient accuracy be given the "approximate" value of $9.05 since it falls in the lowest quarter of the group $9.00 to $9.99.

How precisely a median should be determined depends largely upon the nature of the distribution. The regularity of this series justifies greater nicety in its computation than is typical of most discrete series. Arbitrarily to give it an exact value, however, where it is evident that the differences between the units are clearly unequal, is to allow the *ideal* position of the terms in the group to rob it of much of its significance. This is true only if the median is considered to be more than a mathematical center. It should be interpreted in connection with the kind of series[2] with which it is

[1] In order to have the 113 men distributed throughout this group uniformly and to have the same apply to the groups immediately following and preceding, it would be impossible to assign a man to the last unit of a preceding group and to the first unit of the succeeding group. To do this would result in a concentration at this point. Žižek, in discussing an analogous point, says: "We can distribute 10 values in a class of 200 cents breadth so that the first and the last values coincide with the limiting values of the class; so that the first item coincides with the inferior limit while the last value is as far distant from the superior limit as are the items from each other; or, so that the last item coincides with the superior limit while the first item is as far distant from the inferior limit as are the items from each other. None of these three distributions seems to be free from objection. The first kind of distribution, if carried out in the adjoining classes, would give two items at each class limit. The second and third kinds of distribution do not correspond at all to the postulate of a uniform distribution within the classes. The most correct way of distributing the items uniformly is to assume that they occur at equal intervals even when this distribution is extended to the adjoining classes. To fulfill this condition the first and last of the items belonging to the class must be removed from the class limits to a distance which corresponds to half the magnitude of the interval existing between the items belonging to the class." *Statistical Averages*, pp. 208-209.

[2] In the Dewey *Report on Employees and Wages*, the median is expressed only by group location, and this notwithstanding the fact that the groups are small and the series exceptionally regular.

used. If, in the nature of the case, it can be located with precision, then it should be so located; if otherwise, then it should be given an approximate value.

By extending the principle according to which medians are located, series may be divided into any number of parts. The values of the items dividing a complete series into four equal parts or the halves into two equal parts are called *quartiles*. The dividing position for the lower-half is known as the *first* quartile, or Q1; and of the upper-half, the *third* quartile, or Q3. Obviously, however, these quarter division marks are not averages in the same sense as are the arithmetic mean and the median inasmuch as they have reference to only a part rather than to the whole of a series. Indeed, for their location, the respective parts become complete series. They are not in the same sense typical of, nor may they be considered substitutes for, whole series—an implied characteristic or attribute of an average *per se*.

The first quartile is located with sufficient accuracy by using the formula $\frac{n+1}{4}$, where $n$ is the number of items. The third quartile is located by using $\frac{3(n+1)}{4}$.

But quartiles (quarters), deciles (tenths), percentiles (one hundredths), etc., are not of the nature of averages of the first order; that is, as amounts which may be considered as types or substitutes for detail. Later, in considering the way in which items in series are distributed around their averages, we shall have something more to say about them.[1]

The median and its kindred partition expressions—quartiles, deciles, etc.—are easily located graphically on cumulative curves or ogives by (1) dividing the total measure on the ordinate scale into the required number of parts, (2) extending a line from the point selected parallel to the base or abscissa axis until it meets the ogive, and (3) dropping a perpendicular at this point until it crosses the abscissa scale.

[1] See *infra*, Chapter X, *Dispersion*.

What the scale reading means depends upon the nature of the series. If data are discrete and the perpendicular falls between the measurements, there is no amount which divides the series in half. If the series is continuous in fact and such a condition occurs, then a median amount may be assigned by assuming (1) that another grouping would give such a result, or (2) that another selection of measures expressed in the same way would produce such an amount. If data are grouped and the perpendicular falls within a group, nice interpolation is rarely advisable for discrete although it may be made for continuous series.

An illustration showing the manner in which the median and quartiles are graphically determined in a cumulated *frequency* series is given in Figure 48; the way in which it is done in a cumulated *time* series is shown in Figure 62. In the latter case, the data shown in Table 44 are used.

The first half of the raw cotton imported in the period 1895 to 1913, inclusive, came in between 1895 and approximately September of 1906,[1] that is, during eleven years and eight months. The second half was imported between September, 1906, and the close of 1913, or during seven years and four months. The median period—that is, the half-way period in terms of amounts imported—was September, 1906. In terms of time alone, June, 1904, is the median period. At that time, however, only 40.1 per cent of the total had been imported. These facts are shown graphically on Figure 62. In order to locate the median period in terms of importations, the ordinate axis is bisected at 710,000,000 lbs. and a line extended until it meets the historigram (historical graph) vertically over the period September, 1906. Obviously, in order to locate the median period in terms of time alone, the abscissa axis is bisected at June, 1904, and a perpendicular raised until it meets the historigram horizontally opposite the position 570,000,000 on the ordinate scale.

[1] On the assumption of uniform importation during the year.

TABLE 44

TABLE SHOWING BY YEARS SINGLY AND CUMULATIVELY THE QUANTITY OF RAW COTTON IMPORTED INTO THE UNITED STATES, 1895 TO 1913, INCLUSIVE

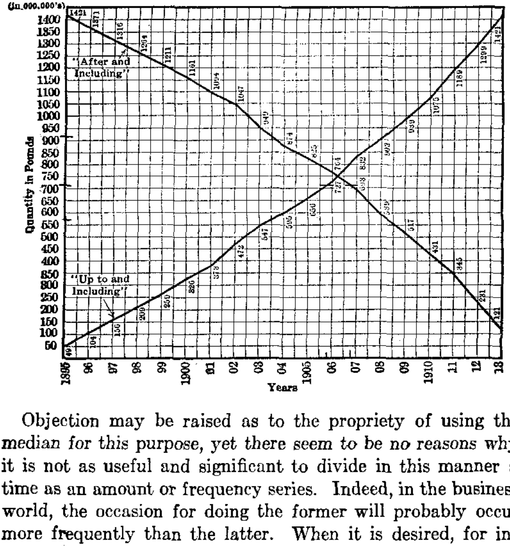
(*Statistical Abstract of the United States*, 1913, p. 669)

| YEAR | AMOUNT OF RAW COTTON IMPORTED, IN POUNDS (*000's* omitted) | | |
|---|---|---|---|
| | NON-CUMULATIVE | CUMULATIVE | |
| | | "Up to and Including" | "After and Including" |
| Total ......... | 1,421,152 | 1,421,152 | 1,421,152 |
| 1895 | 49,332 | 49,332 | 1,421,152 |
| 1896 | 55,350 | 104,682 | 1,371,820 |
| 1897 | 51,899 | 156,581 | 1,316,470 |
| 1898 | 52,660 | 209,241 | 1,264,571 |
| 1899 | 50,158 | 259,399 | 1,211,911 |
| 1900 | 67,398 | 326,797 | 1,161,753 |
| 1901 | 46,631 | 373,428 | 1,094,355 |
| 1902 | 98,716 | 472,144 | 1,047,724 |
| 1903 | 74,874 | 547,018 | 949,008 |
| 1904 | 48,841 | 595,859 | 874,134 |
| 1905 | 60,509 | 656,368 | 825,293 |
| 1906 | 70,964 | 727,332 | 764,784 |
| 1907 | 104,792 | 832,124 | 693,820 |
| 1908 | 71,073 | 903,197 | 589,028 |
| 1909 | 86,518 | 989,715 | 517,955 |
| 1910 | 86,037 | 1,075,752 | 431,437 |
| 1911 | 113,768 | 1,189,520 | 345,400 |
| 1912 | 109,780 | 1,299,300 | 231,632 |
| 1913 | 121,852 | 1,421,152 | 121,852 |

If it is desired graphically to locate the median *amount* in an historical series, amounts and not periods must be arrayed consecutively and each reported performance counted as a frequency of *one*. When this is done, the process is the same as in cumulative frequency series; that is, the amounts cumulated are plotted on the ordinate and the corresponding periods on the abscissa axis.

*FIGURE 62*

CUMULATIVE GRAPHS—HISTORIGRAMS—CONSTRUCTED ON "UP TO AND INCLUDING" AND "AFTER AND INCLUDING" BASES, SHOWING, BY YEARS, IMPORTATIONS OF RAW COTTON INTO THE UNITED STATES




Objection may be raised as to the propriety of using the median for this purpose, yet there seem to be no reasons why it is not as useful and significant to divide in this manner a time as an amount or frequency series. Indeed, in the business world, the occasion for doing the former will probably occur more frequently than the latter. When it is desired, for instance, to distribute expenses over a period, the proportions incurred during one quarter or one half of the time may be of real significance. Of course, amounts, likewise, may be partitioned into equal parts and compared to the time in which incurred. In either case, by plotting the amounts cumulatively and the periods consecutively, the median *positions* may be located and related to each other.

The necessary steps in determining arithmetically the median *amount* imported are given below, and the data arranged as in Table 45. Place the amounts in numerical order and apply the formula $\frac{n+1}{2}$, since $n$ is odd. Thus, $n = 19$. $\frac{n+1}{2} = 10$. The 10th or median item is 70,964,000 lbs. That is, over a period of 19 years the amount imported which stood half-way between the extremes was 70,964,000 and this occurred in the year 1906. The arithmetic mean amount imported is 75,800,000+ lbs. The large items in the latter years *largely explain the difference. In this arrangement,* order of

TABLE 45

TABLE SHOWING DATA OF IMPORTATIONS OF RAW COTTON ARRANGED SO AS TO DETERMINE THE MEDIAN AMOUNT IMPORTED

| PERIODS | FREQUENCIES | IMPORTATIONS IN POUNDS |
|---|---|---|
| Total | 19 | 1,421,152,000 |
| 1901 | 1 | 46,631,000 |
| 1904 | 1 | 48,841,000 |
| 1895 | 1 | 49,332,000 |
| 1899 | 1 | 50,158,000 |
| 1897 | 1 | 51,899,000 |
| 1898 | 1 | 52,660,000 |
| 1896 | 1 | 55,350,000 |
| 1905 | 1 | 60,509,000 |
| 1900 | 1 | 67,398,000 |
| 1906 | 1 | 70,964,000 |
| 1908 | 1 | 71,073,000 |
| 1903 | 1 | 74,874,000 |
| 1910 | 1 | 86,037,000 |
| 1909 | 1 | 86,518,000 |
| 1902 | 1 | 98,716,000 |
| 1907 | 1 | 104,792,000 |
| 1912 | 1 | 109,780,000 |
| 1911 | 1 | 113,768,000 |
| 1913 | 1 | 121,852,000 |

magnitude in the amounts rather than continuity of time is followed. In the former arrangement, the time units are consecutive.[1]

### 3. SUMMARY

The median as an average or summarizing expression should be used with great care. While in its computation all frequencies are required, it is not affected by the size of the items except at or near the middle of a series. This may be a significant weakness when not only the number of times an item appears but also its positive size is important. Theoretically, it is best suited to continuous series or to discrete series in which the measurements are numerous and accurate, and when the scale is small and the groups into which they are merged narrow. It should be considered only as one summary of a distribution, and be compared with the arithmetic mean, and the mode whenever possible.

## V. The Mode

### 1. WHAT THE MODE IS

The mode strictly defined is the value of that item in a series which is most characteristic or common. It is the typical measurement—the one which is found the greatest number of times. But not all series possess a single or even a well-defined mode. Some have more than one mode, while others can scarcely be said to have a mode at all. The mode, therefore, is frequently indefinite, its boundaries being difficult to define, and its position uncertain.

As a form of average, the mode may be used in time, in space, and in condition or frequency series. That which occurs most uniformly during a period of time is modal. For instance, the modal number of calls per day made by a salesman upon his clients is (say) five. Day in and day out, this tends

[1] Respecting the further use of the median in the treatment of time series, see pp. 449-453, *infra*.

to be the most characteristic number. As many calls as ten, and as few as two are exceptions—they are non-modal. Again, the most common daily sail of steamship "X" is 400-450 knots. Under extremely favorable conditions, she has done 600 knots; under adverse conditions, as few as 200. The density of population varies widely from district to district. That condition most commonly encountered is modal; the extremes are not modal. Operating expenses in relation to sales as high as 35 per cent and as low as 12 per cent are occasionally encountered in the retailing of meat. The characteristic or modal rate, however, is in the neighborhood of 20 per cent. Again, most males marry at the ages 25-30, although cases are found where marriage is contracted by youths of 18 and by men of 65. These ages, however, are non-modal—they are not "the rule."

The mode as a statistical short-cut or summary has both a general and a precise usage. In such expressions as those above and in the following it is used to suggest the prevailing condition: "The average man is honest." "The average page contains 300 words." "The average number of words in a line of newspaper type is seven." "The average man takes a '40' coat." "The average length of a class recitation is 50 minutes."

In the second sense, however, it is used more precisely. It refers to a real or to an imaginary measurement found or expected in a series. Where its position is indefinite, frequencies are adjusted by widening the groups into which they fall until a modal group is made to appear.[1] Then within the group, the precise mode is located by interpolation, on the assumption that the frequency of the items in the neighborhood of the mode influences its position in proportion to their respective sizes, or that in a wider universe of which the series in question is but a sample, there is a modal or most frequent measurement.

[1] See Tables 18, 46, 48.

If all measurements were continuous and followed the normal law of error or probability curve, a mode of such precision would no doubt obtain, both in the sample and in the entire "population." But not all series are of this type. Some are discrete, measurements falling at more or less arbitrary units which do not arrange themselves in keeping with the normal curve of error. In such cases, the search for an ideal modal position is illusory. The measurement occurring most times *is* modal, the items appearing above and below it having no influence on its position.

The mode in all cases is a reality—a measurement found either in a series or expected in keeping with some underlying assumption of distribution. But the mode is no less definite—although it is frequently less precise—if in continuous series it is spoken of as falling within certain limits, rather than as being a precise amount.[1] Indeed, where nothing is known as to the manner in which instances in series are distributed throughout a modal group, or about the accuracy of the measurements themselves, a mode which is spoken of as falling within certain limits may be more precise—nearer the truth—than one which is given as a specific amount.

In series which are discrete, the mode generally falls at a particular value. Measurements occur at definite intervals. There is no basis for searching for an ideal mode upon the assumption that the measurements at hand are only approximations, or that a *true* mode would be found if the samples were more numerous. Of course a mode may be made to appear by a manipulation of the frequencies—successively widening the groups into which they fall—but the wider the groups the more unreal does the "mode" as determined in this manner become. Moreover, to interpolate within a group

[1] In a recent study, the writer has defined the area marked by deviations of 20 per cent on either side of the average as modal. See Secrist, Horace, "Expense Levels in Retailing—A Study of the 'Representative Firm' and of 'Bulk Line' Costs in the Distribution of Clothing," *Bureau of Business Research Northwestern University*, Series II, No. 9, Chicago, 1924.

in order to secure a precise mode in such cases is never legitimate because it must be arbitrarily done. It should never be made to appear that there is an exact mode when in fact one does not exist.

The meaning of the mode and the manner in which it is located can be best discussed in connection with concrete cases representing different kinds of series.

## 2. HOW THE MODE IS LOCATED

### *(1) The Location of the Mode in Historical or Time Series*

That which is modal or typical occurs most frequently. The exceptional is not modal. In Table 44, showing importations of raw cotton from 1895-1913, the modal year was not 1913, at which time there was imported almost three times as much cotton as there was in 1901. This is the exceptional year. Years which may be suggested as modal are 1895, 1897, 1898, 1899, 1901, and 1904, in each of which between 45 and 55 million pounds were imported. If the conditions set up to determine the mode be altered so as to include all years in which between 45 and 60 million pounds were imported, 1896 also must be called a modal year, and 55 + millions a modal amount. In this, as in so many cases, the mode is indefinite. The way in which historical series may be treated in order to determine an approximate mode is illustrated in Table 46.

In this table the amounts are arranged in order of magnitude. The grouping is as follows: column 2, 5 million pounds; column 3, 10 million pounds; column 4, 10 million pounds, but starting at 45 million and extending to but not including 55 million; column 5, 15 million pounds; and column 6, 8 million pounds. The amounts are equally common in column 1, no account being taken of the degrees of absolute difference. In column 2 (the grouping being 45 to 50, 50 to 55, etc.) groups 45 to 50, 50 to 55, and 70 to 75 are equally common. By widening them to 10 million pounds, as in column 3, more in-

stances now appear at the group 50-60 million than at any other place. By retaining the 10 million pound group but beginning it at 45 million, a decided concentration appears in the first group. By extending the width to 15 million, the group 45 to 60 shows the greatest concentration, but a secondary mode appears in the group 60 to 75 million. Where is the

TABLE 46

DATA SHOWING IMPORTATIONS OF RAW COTTON INTO THE UNITED STATES, ARRANGED SO AS TO DETERMINE THE MODAL AMOUNT

| Year | Am'ts in 000's | Frequencies | | | | | |
|---|---|---|---|---|---|---|---|
| | | Identical Col. 1 | Approximate, by Groups | | | | |
| | | | 5 Mil. beginning at 45 Mil. Col. 2 | 10 Mil. beginning at 40 Mil. Col. 3 | 10 Mil. beginning at 45 Mil. Col. 4 | 15 Mil. beginning at 45 Mil. Col. 5 | 8 Mil. beginning at 46 Mil. Col. 6 |
| 1901 | 46,631 | 1 | | | | | |
| 1904 | 48,841 | 1 | 3 | 3 | | | |
| 1895 | 49,332 | 1 | | | | | |
| 1899 | 50,158 | 1 | | | 6 | 7 | 6 |
| 1897 | 51,899 | 1 | 3 | 4 | | | |
| 1898 | 52,660 | 1 | | | | | |
| 1896 | 55,350 | 1 | 1 | | | | |
| 1905 | 60,509 | 1 | 1 | 2 | 2 | | 2 |
| 1900 | 67,398 | 1 | 1 | | | | 1 |
| 1906 | 70,964 | 1 | | | | | |
| 1908 | 71,073 | 1 | 3 | 3 | 4 | 5 | 3 |
| 1903 | 74,874 | 1 | | | | | |
| — | — | | | | | | |
| 1910 | 86,037 | 1 | 2 | 2 | 2 | 2 | 2 |
| 1909 | 86,518 | 1 | | | | | |
| — | — | | | | | | |
| 1902 | 98,716 | 1 | 1 | 1 | 2 | 2 | 1 |
| 1907 | 104,792 | 1 | 1 | 2 | | | 2 |
| 1912 | 109,780 | 1 | 1 | | 2 | 2 | |
| 1911 | 113,768 | 1 | 1 | 1 | | | 1 |
| — | — | | | | | | |
| 1913 | 121,852 | 1 | 1 | 1 | 1 | 1 | 1 |

mode? Undoubtedly the most characteristic amount imported when the whole period is considered is less than 60 million pounds. But how much less? The arithmetic mean of the amounts less than 60 million pounds is 50,695,000 and the median 50,158,000. The most characteristic amount with a 10 million group is 46 to 56 million, of which there are seven instances; more narrowly, there are five years in which the amounts imported are between 49 and 56 million. It is probably not wise to locate the mode more accurately than in the group 46 to 54 million (column 6). To do so for this type of distribution would be to strive for too great precision.

While in this case, the modal amount of cotton imported into the United States is probably more accurately stated as falling between 46 and 54 million pounds than by using any precise amount, even these limits are purely arbitrary. Others might with almost equal merit have been chosen.

It should be noted that the amounts in Table 46 are arranged in ascending order, the exact quantities being indicated. The frequencies in this case are the numbers of years in which the amounts imported fall into different sized groups. With any grouping, these must be of uniform size inasmuch as *comparative* frequency is used to secure the mode. An alternative method of presenting the same data would be to set up a series of frequency tables with groups of different widths and to tally opposite each group the number of corresponding cases (years). Of course, if this were done, the *historical* order of the series would be broken just as it is in Table 46. Indeed, for the calculation of the mode, the order of the years is without significance.

If the same data were graphically presented with successive time intervals indicated on the $X$ axis, and the amounts shown as ordinates at the different years, then the typical or modal fact would be indicated by uniformity in the lengths of the ordinates.

When historical data are plotted cumulatively, as in Figure 62, the modal position or positions are shown by the tendency

of the graph to increase or decrease, as the case may be, at a uniform rate. Inasmuch as the chronological order is followed in cumulating, modal amounts will probably not be placed in juxtaposition. If this is so, the dominant characteristic is difficult to locate. The use of the graphic method for determining the mode in historical series is not advocated.

### *(2) The Location of the Mode in Space Series*

Suppose it were desired to find the modal number of passengers carried on different divisions of a railroad; or the modal maintenance cost of road bed for successive miles, data being available respectively by divisions and by miles. The problem would be analogous to that just given concerning imports of cotton for successive years. In the space series, the divisions and miles, respectively, would be the frequencies corresponding to the different numbers of passengers and to total costs. Some sort of grouping would undoubtedly be necessary to determine the modal amount, but the size of the groups would probably have to be arbitrarily selected. Moreover, if the data were graphically presented on the ordinate or $Y$ axis and the successive divisions and miles on the abscissa or $X$ axis, then modality would be indicated by uniformity in the lengths of the ordinates. Similarly, if for successive divisions and miles, the data were cumulated, modality would be shown by the tendency of the graphs to increase or decrease, as the case may be, at a uniform rate. The graphic method, however, is not well suited to determine the mode in such series.

### *(3) The Location of the Mode in Frequency Series*

The measurements of a variable characteristic or attribute of a phenomenon at an instant of time produce what is known as a frequency series. The same type of measurement—as height, for instance—of each member of a class, or repeated measurements of an individual of a class, give such series. Their properties have already been discussed in other con-

nections.[1] We are now interested in the meaning and location of the mode in such series.

Table 17[2] shows the number of real estate mortagages in Wisconsin in 1907, classified by rates of interest. This is a discrete series. The most common interest rate as shown by the table is 5 to 5½ per cent. Of the total number of mortgages—28,961—10,262 had rates falling within these limits. This is the modal *group*, but what is the *mode?* Widening the groups as in columns (b) and (c) of the table produces modal groups at 5 to 6 per cent, and 4½ to 5½ per cent, respectively. The precise mode, however, is in doubt; it is no more accurately approached by the latter process. The truth is that the most common rate is 5 per cent—a conventional unit for borrowed money—and is not revealed by any scheme of grouping.

Moreover, inasmuch as this is a discrete series, there is no reason why one should interpolate for the mode, in an attempt to give effect to the pull which the frequencies adjacent to the modal group might seem to have on the location of the true mode. Instances are not uniformly distributed throughout the modal group, nor through the groups adjacent to it—they congregate on definite units. In this case there is no basis for assuming that the instances are uniformly distributed on either side of a true mode. Accordingly, the smaller the group the better. The mode in this case is not ideally placed at the center of a probability series. The items above and below it do not help to determine its location.

The case, of course, is quite different with continuous series. Tables 18 and 26 and Figure 45 show such series. In these the measurements are only approximations to an ideal, the groupings being arbitrary. A true mode both in the samples and in the complete "universe" may be expected, and it is legitimate on the basis of what is known about the measure-

[1] See *supra*, p. 157 *ff*.
[2] P. 164.

ments to widen the groups until a mode appears. Moreover, its group position once located, it may be more accurately and precisely fixed by interpolation, effect being given to the "pull" of the items adjacent to it. This follows because it is known by hypothesis that if the measurements were more accurately made, and the sample more complete, there would be a true mode. Hence the validity of the attempt to fix it for the series in question.

But statistical series are rarely homogeneous—differences characterize them in other respects than the attribute which is measured. For instance, the carpenters whose wage-rates are measured may differ as to training, kind of work done, etc.; the retail stores whose operating expenses as percentages of sales are compared differ as to size, location, business management, etc. All of these non-homogeneous conditions may make the mode of the aggregate non-typical of the parts. This fact is illustrated in the series in Table 47.

Table 47 shows the number of store-periods (monthly) in retail meat stores in which the ratios of operating expense to sales were classified amounts. For the total, the modal per cent group is 18-20; for stores with annual sales of less than $20,000, it is 20-22; for those with annual sales between $20,000 and $45,000, it is 18-22. For those with annual sales between $45,000 and $75,000 it is 18-20 per cent, and for those with annual sales of $75,000 and over it is 14-16 per cent. What is the mode? In spite of the fact that the modal group is fairly definite for each class of stores and for the total, it varies inversely in size with the amount of business transacted. What is typical for the aggregate is not generally typical of its parts.

In series which are continuous, as are those shown in Table 47, modes may be interpolated for within their respective groups. The manner in which this is done may be illustrated as follows by using the total column in Table 47. The modal group is 18-20 per cent, the number of frequencies being greatest at this point. In the next higher group there are 190 cases,

and in the one immediately below, 170 cases. Combined, these make 360 instances. $\frac{170}{360}$ are exerting an influence to place the mode below the 18-20 per cent group; and $\frac{190}{360}$, to place it above this group. $\frac{170}{360}$ of 2 per cent—the width of the modal group—is 0.94; $\frac{190}{360}$ of 2 per cent is 1.06. Accordingly, the

TABLE 47

NUMBER OF STORE-PERIODS (MONTHLY) IN WHICH RATIOS OF OPERATING EXPENSE TO SALES WERE CLASSIFIED AMOUNTS IN RETAIL MEAT STORES

| TOTAL EXPENSE PER CENT OF SALES | TOTAL (Store-Periods Monthly) | NUMBER OF STORE-PERIODS (monthly) WITH CLASSIFIED YEARLY SALES IN 000'S * | | | |
|---|---|---|---|---|---|
| | | –$20 | $20–$45 | $45–$75 | $75 & over |
| Total | 1088 | 257 | 622 | 143 | 66 |
| 10–12 | 10 | | 8 | | 2 |
| 12–14 | 28 | | 11 | 6 | 11 |
| 14–16 | 108 | 2 | 67 | 17 | 22 |
| 16–18 | 170 | 10 | 110 | 37 | 13 |
| 18–20 | 196 | 19 | 120 | 47 | 10 |
| 20–22 | 190 | 43 | 120 | 20 | 7 |
| 22–24 | 136 | 35 | 89 | 11 | 1 |
| 24–26 | 73 | 31 | 39 | 3 | |
| 26–28 | 54 | 26 | 26 | 2 | |
| 28–30 | 33 | 24 | 9 | | |
| 30–32 | 27 | 18 | 9 | | |
| 32–34 | 14 | 8 | 6 | | |
| 34–36 | 20 | 17 | 3 | | |
| 36–38 | 9 | 7 | 2 | | |
| 38–40 | 11 | 8 | 3 | | |
| 40–42 | 9 | 9 | | | |

* The groups are chosen so as to reflect as accurately as possible one-man, two-man, three-man, four-man and larger stores. This explains the reason for their unequal size.

mode is 18 per cent + 1.06 per cent or 19.06 per cent; or conversely, it is 20 per cent — .94 per cent or 19.06 per cent.

Is there such a mode in reality? What is gained by such nicety of calculation? Is not such an amount pure fiction? Inasmuch as this series is truly continuous, such a mode may in fact appear, yet even in this case too great refinement may have the effect of making the mode unreal. The figures to the right of the decimal point may never be encountered. Yet there is no reason why they may not appear since continuity characterizes the series. There are, moreover, certain advantages in making the mode precise, the chief of which is that in this form it can be compared with the arithmetic mean and median—two other statistical summaries.

But why consider only the frequencies immediately adjacent to the modal group? Why not give weight to all of those below and to all those above this position? There is no reason why this should not be done, but there is little reason for doing it. If a series approaches the normal type, the pull of the items on one side is largely counterbalanced by that of the items on the opposite side. In markedly asymmetrical series only, will the position of the mode be materially changed by giving full effect to the influence of all of the items, and it is precisely these in which a "true" mode is not to be expected.

When frequency series are plotted on a simple graph, the modal position is shown by the maximum ordinate.[1] The meaning of the measurement at this ordinate, however, is different for discrete and for continuous series. How different has already been considered. Such graphic illustrations, in this respect, are unlike those showing time and space series. In the latter, the maximum ordinate shows extreme rather than modal measurements. This follows because at each time or space unit on the $X$ axis, a single instance is illustrated on the ordinate. The mode is shown by ordinates of equal or approximately equal length.

[1] See Figure 45.

On ogives or cumulative graphs of frequency series, the mode or place of greatest frequency appears at the position where the curve passes through the greatest distance vertically in a given distance horizontally, that is, at the position where *the curve is most nearly vertical, or at the point of inflection.* Bowley has suggested the empirical rule of rotating a ruler on the curve at this point in order to determine its exact location. But this method of determining its position is only roughly satisfactory. The modal positions on Figure 48, however, were located in this manner.

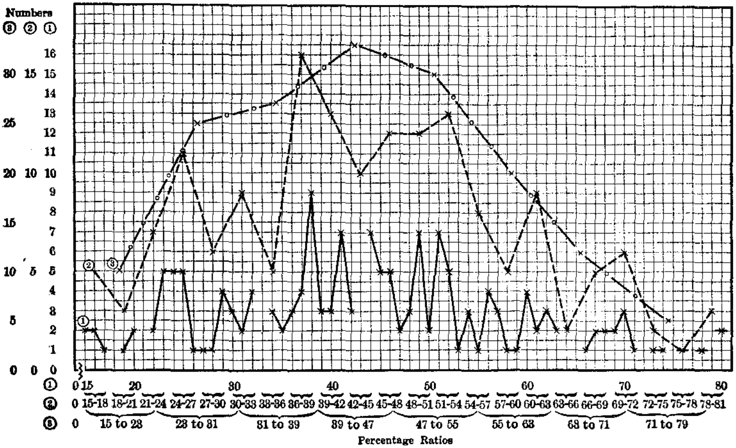
When series are arranged in frequency groups and distributions are irregular, showing no tendency to be dispersed in a definite order around a modal center, it is frequently desirable successively to widen the groups, at the same time altering the frequencies to correspond, until regularity appears. There is always the danger, however, when dealing with discrete series, of concealing the individual peculiarities of the data and of forcing a mode to appear. Group adjustment may be used as a method of correcting a false impression, as, for instance, when data clearly of the continuous type have been distorted from the order which they should properly assume because of the limitations of the units in which they are expressed or by inadequacy of sampling.[1] It is always a question, however, to know how far to carry this synthesizing process.[2] *In effect, it is a method of smoothing and, therefore,* in discrete series, sacrifices individual characteristics in order to secure general impressions. The peculiarities of the whole series dominate those of the parts. It should be remembered that for discrete series, group widening in order to secure regularity of distribution should rarely be employed. This topic was discussed in Chapter VI, and can, therefore, be dis-

[1] See the Table showing the measurements of lengths of lobsters, Chapter VI, p. 165.

[2] See Secrist, Horace, *Readings and Problems in Statistical Methods*, Macmillan, New York, 1920, pp. 278-282, for a discussion by Knibbs, G. H., of "The Theory and Justification of Curve Smoothing."

FIGURE 63

HISTOGRAMS SHOWING THE DISTRIBUTIONS OF RATIOS OF ASSESSED VALUES OF BUILDINGS TO THE ASSESSED VALUES OF LANDS UPON WHICH THEY STAND, NEW YORK CITY, 1914



posed of with this word of caution, and with brief reference to Figure 63.

### 3. SUMMARY

The mode of a statistical series is always represented by actual or implied cases. But not all series have clearly defined modes. Continuous series which by hypothesis follow or approach the ideal distribution of the normal curve may be manipulated in order to secure a true mode. Those which are discrete should not be so treated.

The modes of the parts of an aggregate do not necessarily average or add to the mode of the total.[1] Moreover, this form of a statistical summary rejects all exceptional instances, the type being determined solely by degrees of uniformity. That which is most common is modal. But commonality is frequently difficult to define because so much depends upon the standards by which one chooses to establish it. There can never be a difference of opinion as to the arithmetic mean of a series, but there may be as to the mode. The arithmetic mean is rigidly defined; but a mode is not.

## VI. The Geometric Mean

The geometric mean of the values of the items in a series is the *nth* root of their product. Rather than *adding* the values together and *dividing* their sum by the number of items, as is done in calculating the arithmetic mean, the geometric mean is secured by multiplying the values of the items together and taking the root corresponding to the number of items. The formula is: Geometric Mean $= \sqrt[n]{p_1 \times p_2 \times p_3 \times \text{——} p_n}$; $p_1$, $p_2$, $p_3$ —$p_n$ referring to the values of the different items, and $n$ to the number of items. The

[1] The Bureau of Business Research, Harvard University, adjusts the modes of the different expenses in conducting retail and wholesale stores so as to add to the modes of the total expenses. This practice is equivalent *rigidly* to defining the mode, a practice to be justified only when distribhntions are of the probability type.

arithmetic mean of 2, 3, and 4 is 3; the geometric mean is $\sqrt[3]{2 \times 3 \times 4} = 2.9$ (approximately).

The geometric mean is most easily calculated not by successively multiplying a series of numbers together and extracting the corresponding root, but by using logarithms. Certain rules for their use are as follows:

(1) To multiply a series of numbers together add their logarithms. The natural number corresponding to the result is equal to the product of the numbers.

(2) To divide one number by another, subtract the logarithm of the divisor from the logarithm of the dividend. The natural number corresponding to the result is the quotient.

(3) To raise a number to any power, multiply the logarithm of the number by the power exponent. The natural number corresponding to the product is the required power of the number.

(4) To extract any root of a number, divide the logarithm of the number by the index of the root. The natural number corresponding to the quotient is the root of the number.

It is desired, for instance, to compute the geometric mean of the ratios of total operating expenses to sales for all stores as shown in Table 47. The method is as follows:

(1) Find the log of 11—the middle of the first group. This is 1.0414. Raise this to the 10th power, that is, the power corresponding to the frequency. This is done by multiplying the log 1.0414 by 10 which gives 10.4140.

(2) Find the logs of the centers of each of the other groups, and multiply them respectively by the powers or the corresponding frequencies.

(3) Add the products as found in (1) and (2) above.

(4) Divide the total by the number of powers, that is, by 1088.

(5) Find the natural number corresponding to this quotient. This is the required geometric mean.

Each of the steps through which the above data must be

carried in order to calculate the geometric mean is shown in Table 48. The geometric mean ratio is 20.7, the arithmetic mean 21.3, the median 20.4, and the mode 19.1.

TABLE 48

TABLE SHOWING THE STEPS USED IN CALCULATING A GEOMETRIC MEAN

| RATIO EXPENSE TO SALES (Center of Group) | LOGS | POWERS | PRODUCTS OF LOGS AND POWERS |
|---|---|---|---|
| 11 | 1.0414 | 10 | 10.4140 |
| 13 | 1.1139 | 28 | 31.1892 |
| 15 | 1.1761 | 108 | 127.0188 |
| 17 | 1.2304 | 170 | 209.1680 |
| 19 | 1.2788 | 196 | 250.6448 |
| 21 | 1.3222 | 190 | 251.2180 |
| 23 | 1.3617 | 136 | 185.1912 |
| 25 | 1.3979 | 73 | 102.0467 |
| 27 | 1.4314 | 54 | 77.2956 |
| 29 | 1.4624 | 33 | 48.2592 |
| 31 | 1.4914 | 27 | 40.2678 |
| 33 | 1.5185 | 14 | 21.2590 |
| 35 | 1.5441 | 20 | 30.8820 |
| 37 | 1.5682 | 9 | 14.1138 |
| 39 | 1.5911 | 11 | 17.5021 |
| 41 | 1.6128 | 9 | 14.5152 |
| Total............. | | 1088 | 1430.9854 |

Log 1430.9854 ÷ 1088 = 1.3152 Log
The natural number of Log 1.3152 = 20.7 (approximately).
This is the geometric mean.

But such a use would rarely be made of this average. This example is inserted so as to show the manner in which the computation is made. More appropriate uses of this average are discussed below in Chapters XV and XVI.[1]

[1] *Passim.*

### VII. The Properties of the Arithmetic Mean, the Median, the Mode, and the Geometric Mean Compared and Contrasted

The properties of the different averages discussed in this chapter when computed from statistical series may be summarized as follows:

| *Characteristics or Properties* | *Averages Represented* |
| --- | --- |
| 1. Data Required | |
| (1) All the frequencies and the exact size of all amounts. | Arithmetic Mean, Geometric Mean |
| (2) All the frequencies but the exact size of only certain amounts. | Median |
| (3) Only certain frequencies and certain amounts. | Mode |
| 2. Representation in a Series | |
| (1) May be represented | Arithmetic Mean, Median, Mode, Geometric Mean |
| (2) Must be represented (actually or ideally) | Mode |
| 3. Order of Arrangement for Calculation | |
| (1) A definite order | Median and Mode |
| (2) Any order | Arithmetic Mean<br>Geometric Mean |
| 4. Influence of Extreme Items | |
| (1) Proportional to their size and frequency | Arithmetic Mean |
| (2) Proportional to frequency alone | Median |
| (3) Small numbers given proportionally larger influence | Geometric Mean |
| (4) No influence | Mode |
| 5. Relative Size in the Same Series | |
| (1) Permanent differences | Arithmetic Mean exceeds the Geometric Mean in all cases except when all values of a series are equal to each other |

| | |
|---|---|
| (2) Variable differences | Relative size of Arithmetic Mean, Median, and Mode depends on the distribution of the items in series |
| 6. Relative Position in Series | Median always lies between the Arithmetic Mean and Mode in mono-modal distributions |
| 7. Degree of Precision of Measurement | |
| (1) Definite | Arithmetic Mean and Geometric Mean |
| (2) Often Indefinite | Median and Mode |
| 8. May be Interpolated for | Median and Mode |
| 9. May be Located Graphically | Median and Mode |
| 10. Can be Algebraically Treated | Arithmetic Mean and Geometric Mean |
| 11. From Averages of the Parts, Averages of a Total may be Secured | Arithmetic Mean and Geometric Mean |
| 12. When Substituted for Each of the Original Items | |
| (1) Sum of, remains the same | Arithmetic Mean |
| (2) Product of, remains the same | Geometric Mean |
| 13. Sum of the Deviations from, a minimum | Median |
| 14. Algebraic Sum of the Deviations from, Equals Zero | Arithmetic Mean |
| 15. Can be Calculated from Totals Only | Arithmetic Mean (isolated) |

## VIII. The Average to Use—Some Typical Cases where Choice is Important [1]

Suppose a firm were interested in the experience of one of its salesmen as a basis for promotion to a new territory or

[1] Examples in which it is desirable to use the geometric mean are given in Chapters XV and XVI.

to an advanced wage or salary scale. It is further supposed that the sales record of this man is available over an extended period, the sales being listed by territory, by grade of commodity, by prices of the article sold, by profits realized by the firm, by the length of time utilized in making them, by cost to the firm in present salary and expenses, etc. Can the sales experience of this man be averaged? If so, what average shall be used? Is the arithmetic mean—an average of sales during good and bad days, of sales among all classes of buyers, of those requiring one call and those requiring close following up, of small and large sales, of those upon which small as well as large profits are realized, etc.—a suitable measure of a salesman's activity?

If it is not, then probably a weighted average would be more appropriate, especial importance being given to large sales, sales of goods upon which a high rate of profit is made, etc. Is an average which takes account of the bad days and the small sales, of the good days and the large sales, but which gives no more importance to one of them than to another more satisfactory for this purpose? Such a line of thought suggests the advisability of using the median. But, comes the retort from one who approaches the problem from another point of view: "This man has had a consistent record of a high order, and it is neither fair to the man nor to the company to give weight to his misfortunes. The facts show that he can be expected to make such and such a record—the overwhelming percentage of his sales are of this character; or, in other words, the percentage of the time in which he fell below a high standard is negligible and should be given no weight. If his mistakes and failures are considered, a premium will be put upon mediocrity and insufficient recognition given to real merit." Such an argument suggests the wisdom of using the mode.

It may be argued that it is unwise to let any one set of circumstances govern, no matter from what angle the problem

is approached, and, undoubtedly, this is true. However, no matter how carefully the promotion is considered, if the facts above indicated are held to be germane, it is necessary to decide upon the weight to be assigned to the approaches indicated in these different averages. It is, of course, conceivable that the various averages would not be materially different. If this is true, any one of them may be used. As to whether averages can be used is one question: which one to use, in case they are allowable, is quite another. It is the latter question which is now being discussed.

Again, suppose that one were interested in the time necessary to reach his work—a fact governing his location for residential purposes—and that there existed but one available means of transportation. Is it the arithmetic mean time, the median time, or the modal time in which the distance is traveled which is of interest? Delays happen even in connection with the best transportation service.[1] Should the possibility of these be considered or should they be regarded as negligible on the ground that they are irregular and uncertain? If one sets great weight upon punctuality, he undoubtedly will allow for this factor in spite of its contingency.

On the other hand, if the transportation company in question were advertising its service, it would feature the typical or modal if not the shortest performance. If many measurements were taken of the required time to make the trip, it is doubtful whether the differences between the various averages would be large. The distribution of frequencies would tend to conform to the normal law of error curve and the averages closely to agree. On the other hand, if few measurements were taken, and if the delays were frequent, the characteristic or modal might be widely different from the mean time. There would be no tendency for delays to be compensated for by

[1] See "Report" of the *Chicago Traction Subway Commission*, "On a United System of Surface, Elevated and Subway Lines," pp. 272-274, Chicago, 1916, for an analysis of the classified causes of one year's reported delays of more than five minutes' duration on the surface lines.

*exceptionally quick service, since most of the runs would be* made according to schedule. The arithmetic mean would exceed both the median and the mode. It is precisely this fact which needs to be considered by the person who desires to reach his office each morning at or before a stated time, and which the advertising manager of the company desires not to bring to the attention of the public. It is evident that the averages accurately reflect the characteristics of the data, but they call attention to different things.

One might be interested in the "average" suit of ready-made clothes turned out by a clothing concern, but the kind of an average best suited to his purposes will depend upon what those purposes are. If he is in the production side of the business his interest is in typical or standard sizes determined for him by the physical facts of size and proportion of men. The great majority of sales will be to individuals who conform within narrow limits to standard measurements. The manufacture of these garments constitutes his problem. His interest lies in the modal suit; not in the median nor in the arithmetic mean, as such. If he considered the arithmetic mean and manufactured his garments according to the sizes determined by such a calculation, it is doubtful if his customers could be fitted, since such measurements imply that the exceptionally large and the exceptionally small will affect the measurements of suits designed for the great homogeneous and standard majority. If large quantities of suits were manufactured, it is true that the mode, the median, and the arithmetic mean sizes would closely agree; but by the prudent producer this agreement would be taken for granted only where production was on the largest scale.

Likewise, if the value instead of the size of the "average" suit were uppermost in one's mind, it is doubtful if the arithmetic mean would be particularly enlightening. Such a figure is too general, too indefinite, for any but the most superficial purposes. Some sizes tend to be normal; this

grows out of a physical fact. Values tend to be normal or characteristic too, but their normality is not reflected in an arithmetic mean, as it is in the case of sizes, since all values may or may not be represented in the various sizes manufactured. Suits which can be manufactured according to set measures and in large quantities, other things being equal, tend to be cheap. Suits which are manufactured only to special order and in relatively small quantities, other things being equal, tend to be dear. The exceptional in either case would be weighted heavily and the characteristic be far different from the mean price. As a basis for roughly estimating profit an arithmetic mean price may be all that is required, but for shaping a selling policy an intimate study of the characteristic prices for the various types of demand is necessary. This is merely another way of saying that only homogeneous data can be properly averaged, and that the merits of each average must be settled in the light of its use.

The errors into which one may be led by indiscriminately using an average of non-homogeneous data are admirably shown in Table 49 giving deaths and death-rates of married and unmarried men in Scotland.[1]

"The first striking fact which this table reveals is that the death-rate of the bachelors was double that of the married men between the ages of 20 and 25. As its persons became older, this excessive difference in the death-rates of the married and the unmarried decreased slowly and regularly, showing the difference in favor of the married men at every period of life. It is thus proved that the state of bachelorhood is more destructive to life than the most unwholesome trades. When we come to the total death-rate at all ages, however, the very reverse is the case. The general death-rate among married men is very much higher than that among single men; so that, while only 1,723 bachelors died during the year out of every 100,000 bachelors, 2,338 married men died out of a like number of married men.

[1] See also an analogous case in Secrist, Horace, "A Statistical Paradox," *Journal of the American Statistical Association*, June, 1923, pp. 776-780.

"This apparent contradiction may be explained as due to the fact that the number of bachelors being far greatest at that period of life when the mortality is very low, namely, from 20 to 24, whereas the number of married men is greatest at those periods of life when mortality is high, seeing that mortality increases with age.

TABLE 49

TABLE SHOWING DEATHS AND DEATH-RATES OF MARRIED AND UNMARRIED MEN IN SCOTLAND, 1863, CLASSIFIED BY AGE GROUPS

(From the 9th Detailed Report of Dr. James Stark to the Registrar-General of Births, Deaths, and Marriages in Scotland)

| Ages | Married | | | Unmarried | | |
|---|---|---|---|---|---|---|
| | Number Living | Deaths | Death-Rate | Number Living | Deaths | Death-Rate |
| All ages | 503,376 | 11,765 | 23.4 | 243,259* | 4,189 | 17.2 |
| 20–25 | 22,946 | 137 | 6.0 | 106,587 | 1,251 | 11.7 |
| 25–30 | 54,221 | 469 | 8.7 | 48,618 | 666 | 13.7 |
| 30–35 | 66,153 | 600 | 9.1 | 25,962 | 383 | 14.8 |
| 35–40 | 63,858 | 690 | 10.8 | 15,857 | 253 | 16.0 |
| 40–45 | 62,645 | 782 | 12.5 | 12,311 | 208 | 16.9 |
| 45–50 | 54,505 | 869 | 15.9 | 8,824 | 179 | 20.3 |
| 50–55 | 49,591 | 880 | 17.7 | 7,636 | 205 | 26.8 |
| 55–60 | 38,006 | 929 | 24.4 | 5,550 | 142 | 25.6 |
| 60–65 | 35,920 | 1,216 | 33.9 | 5,242 | 227 | 43.3 |
| 65–70 | 22,021 | 1,134 | 51.5 | 2,848 | 156 | 54.8 |
| 70–75 | 16,029 | 1,291 | 80.6 | 2,021 | 205 | 101.4 |
| 75–80 | 9,716 | 1,135 | 116.8 | 1,081 | 157 | 145.4 |
| 80–85 | 5,477 | 953 | 174.0 | 513 | 101 | 196.9 |
| 85–90 | 1,708 | 488 | 285.7 | 151 | 32 | 211.9 |
| 90–95 | 449 | 137 | 305.1 | 50 | 21 | 420.0 |
| 95–100 | 103 | 40 | 388.4 | 6 | 3 | 500.0 |
| 100 and above | 28 | 15 | 535.7 | 3 | | |

* As reported. The correct total from the addition is 243,260. The table is quoted from Bliss, George I.—"The Influence of Marriage on the Death-rate of Men and Women," in *Quarterly Publications of the American Statistical Association*, March, 1914, p. 55.

Furthermore, almost half of all the deaths of the bachelors occur before the thirtieth anniversary, at which period the mortality is much lower than at the more advanced periods of life. When the whole deaths at all ages are thrown together and compared with the total bachelors living, the general mortality seems to be little higher than that due to the earlier period of life. Among the married men, on the other hand, the greatest number of deaths occur between the sixtieth and seventy-fifth year of life, at which period the mortality is high as compared with the number living. Consequently, when the total deaths of husbands of all ages are compared with the total living, a high mortality seems to have prevailed, because the persons were all so much older when they died than were the bachelors. Therefore, comparing the total deaths of the married at all ages with the total deaths of the bachelors, necessarily leads to a false conclusion. In comparing mortality rates of two or more classes, to be correct, it must be limited to comparing at each age group, and the smaller we take the age group the more nearly correct are the rates."[1]

While this illustration is drawn from mortality statistics, and seems to have little or no bearing on the problems of the business man, except in so far as it illustrates the error into which one may be led by making his basis of generalization too broad, and therefore his conclusion too indefinite, it suggests a problem of practical import to the business world.

In most states, laws now require that employers of labor provide in some manner for the compensation of accidents which occur to their employes while engaged in the regular course of business. Because of the failure to define an "accident," and because accidents which occur are related to so broad a base, without differentiating between hazardous and non-hazardous occupations, slight and severe accidents; and because of the failure to keep accurate records of accidents, employers have not had until recently, if they now have, an adequate basis for computing accident risks.[2]

[1] *Quarterly Publications of the American Statistical Association,* March, 1914, p. 56.

[2] Rubinow, I. M., "The Standard Accident Table as a Basis for Compensation Rates," *Quarterly Publications of the American Statistical Association,* March, 1915, pp. 358-415.

Discrimination between severe and minor accidents, and hazardous and non-hazardous conditions of employment, is the first essential to clear thinking about accidents, and a partial guaranty of the reasonableness of insurance premiums.[1] A rough arithmetic mean, a median, or a mode, *per se*, is not enough. What is necessary is the determination of the characteristic accident rate, not for industries as a group, but for conditions of employment, definitely standardized, within each industry.

Statistics should always relate to definite conditions and circumstances. Duplicate these and the statistical facts are likely to be repeated. Alter them and the consequences are different. Before a policy can be mapped out on the basis of statistical facts alone, or given consequences said to follow from given conditions, the latter must be definitely and clearly defined and their boundaries indicated.

So-called statistical laws operate with implacable regularity only when conditions producing them occur with unchanging persistence. To establish beyond cavil cause and effect requires not only that statistical data be referred solely to the conditions that produce them, but also that the statistical means employed to interpret them be appropriate to the purposes in mind. There is no excuse for assigning meaning to averages without taking the trouble to determine the conditions which produce them or their suitability to the cases in point.

"An average is not to be regarded as a secret something which determines events. This blunder is often made in social statistics. After finding a certain average in human affairs, we conclude that some secret fate is at work. By the aid of a little rhetoric we easily persuade ourselves that an event is fully accounted for when 'the law of averages' demands it. 'There may be an average in birth and death and crime, but, after all, the average is not responsible for any of them. It takes something more potent than an average to produce typhoid fever or to crack a safe.'"[2]

[1] *Ibid.*, pp. 358 ff.
[2] Coffey, P., *The Science of Logic*, Longmans, London, 1912, Vol. II, p. 291.

To employ an average suggests the formulation of a judgment or a conclusion following from a full consideration of detail which it replaces. An average represents the culmination of a process of thought, which when removed from the steps required for its determination is likely to be assigned new meanings and used for purposes foreign to those for which it was designed. Given statistical application, this means that chronologically averages come late in the process of analysis. They should be used with discrimination and supported by detail, with the realization that they emphasize the generalizations and comparisons which seem to be warranted after a careful and painstaking scrutiny of the problem from the angle from which it is approached.[1]

The functions of averages are unmistakable; the justification of employing them must be determined by an appeal to all the facts and in the light of the peculiarities characteristic of the different types. As a statistical caution let it be said: *Do not rush headlong into the use of averages. They are commonly but vaguely understood, and it is the particular function of the statistician to adopt that caution and circumspection in the use of numerical facts which the seeming exact-*

[1] "But however often an average may have been confirmed, we can never attribute to it the importance of being by itself the expression of any necessity. Every result is necessary when its conditions are given; every particular instance was necessary in so far as from the given conditions it could only be such and no other; all individual determinations and differences in the particular cases, which were neglected by the average, were necessary; the most extreme deviations were necessary, and it will also be necessary, if all the particular conditions recur in exactly the same way, that they should again have the same results, and that therefore the sum of the results will be the same. . . .

"Such uniformities of numbers and averages are primarily mere descriptions of facts which need explanation as much as the uniformity of the alternation between day and night; and the explanation can be found only where the actual conditions . . . are forthcoming. But these are the concrete conditions of the particular instances counted, they are not directly causes of the numbers; it is only the nature of the concrete causes which can show it to be necessary for the effects to appear in certain numbers and numerical relations." Sigwart, C., *Logic*, Swann Sonnenschein Co., London, 1895, Vol. II, p. 490.

*ness of his tools appears not only to suggest but to make imperative.*

## IX. Summary and Conclusion

(An average should be considered as derivative and as summarizing and characterizing data in a single expression.[1] The average best suited for a particular use depends upon the purpose one has in mind. Frequently, it is desirable and necessary to compute not only the arithmetic mean but also the median and mode in order to safeguard oneself against criticism and to reflect types of distributions more in detail. The relations of these averages one to the other are interesting. If it is remembered (1) that the computation of the arithmetic mean and the median requires all the frequencies; (2) that the former is affected by both the size of items and frequencies, while the latter is affected by frequencies and not by the size of items except those at or near the middle; and (3) that in the computation of the mode both the size and frequencies of exceptional items are ignored, then it is evident that in changing the order or number of frequencies the mode is scarcely affected at all; the median is only slightly affected, and the arithmetic mean violently affected.

No single average suffices for all purposes. Each is affected differently by arrangement, frequency, and size of items, and should be used with a full knowledge of the peculiarities of distributions. One is never justified in employing a short-cut expression in order to describe a complex whole unless he re-

[1] An average "is an abbreviation, and it has so much in common with the ordinary logical abstract concept that it neglects all differences, and we cannot tell from it how far the numbers from which it is obtained, or which it has to represent, may differ from each other. It is, however, inferior to the general concept in so far as the latter is a statement of what is the same in all the particular instances, while the average is merely a fictitious value which may never actually occur in any particular case, and which by itself does not even justify us in expecting that the majority of the particular instances in a region will approximate to it." Sigwart, C., *Logic*. Vol. II. p. 487.

alizes what its use implies. Too frequently averages are used without discrimination. Derivative expressions of this character are often imperfect substitutes for detail. Frequently, an exceptional instance which would be ignored in the use of the mode is that particular instance in which one has greatest interest. On the other hand, the inclusion of an exceptional item in determining the arithmetic mean may serve to so prejudice it as to give a wholly erroneous picture of the characteristics which are dominant. The average to be used is invariably a function of the purpose which one has in mind.

As classified data are more readily understood and compared than those in heterogeneous form, and tabular arrangement superior to unscientific classification, so summary expressions of complex data in the form of averages are frequently more significant than the detail. The passage, however, from the particular to the general—that is, from details to averages—offers precisely the opportunity for eliminating the peculiar and significant features of discrete series. In the case of continuous series the conditions are somewhat different. As the widening of groups may result in a more accurate expression of a general tendency or an ideal distribution, so a more accurate expression of a complex whole may result from the use of a single unit, as mean, median, or mode.

Caution, foresight, and analysis are necessary at every step in the use of averages—caution as to the averages to be employed, foresight as to the meaning which may be attached to them, and analysis as to the possibilities of data being characterized in such a manner. The following tests should always be applied: Is it possible to employ a single expression to depict the details which are essential in order to view the data in all their bearings? Is the greatest interest in the characteristic feature, in the median position, or in the center of gravity at which the arithmetic mean falls? Is it necessary to employ all of these descriptive units? No single answer to these various inquiries can be given. The use of an average

may be legitimate and still the question as to the most appropriate average be left in doubt. One cannot answer the first question, as it were, by intuition. Data must be analyzed and the functions of averages in general and in particular be clearly understood before answer can be given. As caution and analysis are necessary in the employment of averages, so discrimination and judgment are necessary in assigning importance to them when used by others.

A fitting close to the discussion of averages is found in the words of Dr. John Venn. "Every sort of average—and there are many such sorts—is a single fictitious substitute of our own for the plurality of actual values existent in the results which are naturally or artificially set before us. It is impossible, therefore, for the former, in any case, effectually to take the place of the latter. But the extent to which it may succeed or fail in doing so will depend upon the nature of the facts presented to us, and still more upon the precise object we have in view."[1]

### References

Bowley, A. L., *An Elementary Manual of Statistics,* MacDonald & Evans, London, 1915, Chapters III and IV, pp. 15-35 (especially Chapter III).

Bowley, A. L., *Elements of Statistics,* 4th Edition, King, London, 1920, Chapter V, pp. 82-109.

Elderton, W. P. and E. M., *Primer of Statistics,* Black, London, 1910, Chapters I and II, pp. 1-23.

Fisher, Irving, *The Making of Index Numbers,* Houghton Mifflin, Boston, 1922, Chapter II, pp. 11-43.

Jones, D. Caradog, *A First Course in Statistics,* Bell, London, 1921, Chapters IV and V, pp. 22-41.

Kelley, Truman L., *Statistical Method,* Macmillan & Co., New York, 1923, Chapter III, pp. 44-69.

King, W. I., *Elements of Statistical Method,* Macmillan, 1912, Chapter XII, pp. 121-140.

Mills, Frederick C., *Statistical Methods Applied to Economics and Business,* Holt, New York, 1924, Chapter IV, pp. 97-146.

[1] Venn, Dr. John, "On the Nature and Use of Averages," *Journal of the Royal Statistical Society,* Vol. LIV, 1891, p. 447.

RUGG, HAROLD O., *Statistical Methods Applied to Education*, Houghton Mifflin, Boston, 1917, Chapter V, pp. 97-148.

SECRIST, H., "The Use of Averages in Expressing the Wages and Hours of Milwaukee Street Car Trainmen," *Publications American Statistical Association*, Volume 13, pp. 279-298 (1912).

YULE, G. UDNY, *Introduction to the Theory of Statistics*, Griffin, London, 1911, Chapter VII, pp. 106-132.

VENN, J., "On the Nature and Use of Averages," *Journal of the Royal Statistical Society*, Volume LIV, 1891, pp. 429-448.

ŽIŽEK, FRANZ, *Statistical Averages* (translated by W. M. Persons), Holt, New York, 1913, Part I, "Nature and Purposes of Averages," Chapter VI, pp. 92-127; Part II, "The Arithmetic Mean," Chapter II, pp. 138-163; Part II, "The Median," Chapter IV, pp. 199-221; Part II, "The Mode," Chapter V, pp. 222-247.

# CHAPTER X

## DISPERSION

### I. Introduction

The preceding chapter was concerned with averages of the "first order"—those statistical summaries computed from the gross items in different kinds of series. It was learned that they have different properties; that they require the details from which they are calculated to be treated differently; that some ignore or treat lightly exceptional instances, while others attach to them marked significance; etc. Notwithstanding their differences, however, they all have one common purpose —that is, to serve as substitutes for or types of the detail which they replace.

But different averages may and generally do give different "types" for the same series. Which, then, is to be selected? The answer to this question must be determined in the light of the purpose which one wants the type to serve. As the purpose differs, the selection of the averages must of necessity change.

But averages of the "first order" while useful never fully characterize the detail from which they are made up. In all but the rarest cases some or all of the items differ from the one or ones which are selected as a type. Some measure of the differences from the average is necessary. Averages of the "second order" serve this purpose. By their use a type not of the gross items but of the differences of these from some center or position is secured. Indeed, in some cases, more than a type is required. To average them is equivalent to doing for them the same thing which is done for the gross

items—that is, merging their dissimilarities (by using an arithmetic mean); selecting those which are typical (by using a mode); or choosing the one centrally located (by using the median). An alternative to the selection of a type is to employ some form of a distribution of detail, but this is often unsatisfactory for the same reason that it is when dealing with the gross items themselves. Precise summaries are needed if for no other reason than because of their brevity.

The things about statistical series which it is desirable to know are: (1) the number of instances involved; (2) the average, central or typical fact; (3) measurements of the differences of the individual items from each other or from their averages; and (4) summaries of the manner in which the items are distributed about their average. To secure the summaries in (2), averages of various sorts are computed; to obtain those in (3), measures and coefficients (ratios) of dispersion are calculated; and to get those in (4), measures and coefficients (ratios) of skewness (degrees of asymmetry) are determined. Summaries of the second type are discussed in the preceding chapter; those of the third type, in this; and those of the fourth, in Chapter XII.

## II. The Meaning of Dispersion

In statistics, there are two uses of the term "dispersion." One is general, calling attention to the fact that the items in statistical series differ in size. A wage series with items running from $4.00 to $12.00 per day is spoken of as having greater dispersion than another one having items ranging from $5.00 to $8.00 per day. That is, the instances are dispersed or scattered over a wider range in the first than in the second case. The amplitude of variation is greater. In a more precise sense, the term is used as an absolute or relative measure of the differences of the items in a series from the average or characteristic amount. The first use calls attention to the limits within which data fall; the second use, to an

amount (absolute or relative) by which the data differ from a selected standard or type. The two uses are fundamentally different. In what way will appear as the methods of measuring dispersion are described.

## III. Measures and Coefficients of Dispersion

### 1. THE METHOD OF LIMITS

Dispersion in the general sense indicated above is shown by the "method of limits," the complete range of the values of the items or other conventional divisions being used for this purpose. Examples of the use of different limits, and of ways of stating the degrees of dispersion will illustrate this method.

#### (1) *The Range*

The simplest way of expressing the degree of difference between items in statistical series is to choose the extreme limits within which they fall—that is, to select a minimum and maximum above and below which all items are found. In frequency series, however, it is difficult to define the limits exactly if the groups at the ends of the series are open. When this occurs, approximation is necessary. In historical series, on the other hand, approximation is unnecessary—the actual amounts always being given. Moreover, the selection of extremes in the two kinds of series has a different significance. In those of the frequency type the extreme measurements are relatively few. This is always the case in series which are symmetrical and in those which approach the normal curve of error form.[1] Accordingly, to select the extremes is to choose non-typical cases. In historical series, on the other hand, since there is no presumption of normal distribution, either extreme may be as nearly typical as any other measure. But in either case, to measure dispersion by the range gives no idea of the

[1] For an illustration of the ideal curve of error, see p. 378.

distribution between the extremes. Illustrations will show the force of this contention in typical cases.

In the historical series in Table 45, the extremes are 46,631,000 lbs. and 121,852,000 lbs. This fact carries a certain amount of significance but it does not indicate the dispersion of the items between these limits. It does, of course, rule out such ideas that as small and as large amounts, respectively, as 20,000,000 lbs. and 200,000,000 lbs., for instance, were imported. It does not, however, indicate the fact that the minimum amount is far more characteristic of the series than is the maximum. Moreover, the extremes might remain the same and the distribution between them be quite different.

In the frequency distribution in Table 43 the limits are $5.00 and $14.99, but such amounts are exceptional. Moreover, the frequencies in the lowest group, $5.00 to $5.99, are fifteen times as numerous as those in the highest group, $14.00 to $14.99. As to the distribution of values between these limits, the range tells us nothing. Something more than this

TABLE 50

TABLE ILLUSTRATING THE CUMULATIVE- OR MOVING-RANGE METHOD OF SHOWING DISPERSION IN HISTORICAL SERIES

| Years | Importations | |
|---|---|---|
| | Amounts in (000's) lbs. | Per cent |
| 1895 to 1913 | 1,421,152 | 100.0 |
| 1895 to 1900 | 326,797 | 23.0 |
| 1895 to 1905 | 656,368 | 46.2 |
| 1895 to 1910 | 1,075,752 | 75.7 |

The data may be put in this manner:

| | | |
|---|---|---|
| 1895 to 1913 | 1,421,152 | 100.0 |
| 1910 to 1913 | 431,437 | 30.4 |
| 1905 to 1913 | 825,293 | 58.1 |
| 1900 to 1913 | 1,161,753 | 81.7 |

crude measure is necessary. This "something" is supplied by the cumulative- or moving-range method described below.

If the time series is used, some such dispersion summary as that shown in Table 50 may be prepared, the amount of detail being varied to suit the needs of the problem.

Applying the same method to the frequency series in Table 13, p. 287, an arrangement similar to that in Table 51 might be used.

TABLE 51

TABLE ILLUSTRATING THE CUMULATIVE- OR MOVING-RANGE METHOD OF SHOWING DISPERSION IN FREQUENCY SERIES

| Amounts | Frequencies | |
|---|---|---|
| | Amounts | Per cents |
| As much as $5 but less than $15.00 . . . | 434 | 100.0 |
| As much as $5 but less than $ 8.00 . . . | 121 | 27.9 |
| As much as $5 but less than $11.00 . . . | 374 | 86.2 |
| As much as $5 but less than $14.00 . . . | 433 | 99.8 |
| Or in this manner | | |
| Less than $15 but more than $ 4.99 . . . | 434 | 100.0 |
| Less than $15 but more than $13.99 . . . | 1 | .2 |
| Less than $15 but more than $10.99 . . . | 60 | 13.8 |
| Less than $15 but more than $ 7.99 . . . | 313 | 72.1 |

The method of showing dispersion by the cumulative- or moving-range consists in establishing a series of cumulations by adjusting the sizes of groups. Grouping may be begun from either end and carried forward step by step. The thing that is striven for is a summary which characterizes the complete distribution.

But the use of the range method whether stationary or moving does not make it possible to compare the *relative* dispersion of two series expressed in different units. Such a comparison can be made, however, by reducing the absolute measures to relative bases. This may be done by dividing

the difference between the extremes by their sum. In the cases used for illustration, the coefficients or ratios of dispersion are as follows:

$$\text{In the historical series: } \frac{121{,}852{,}000 \text{ lbs.} - 46{,}631{,}000 \text{ lbs.}}{121{,}852{,}000 \text{ lbs.} + 46{,}631{,}000 \text{ lbs.}} = .45$$

$$\text{In the frequency series: } \frac{\$15 - \$5}{\$15 + \$5} = .50$$

But to show dispersion, limits other than the extremes may be selected. The 1st and 9th deciles are often used for this purpose. The measure of dispersion based upon them is secured by taking their difference, and the coefficient obtained by dividing this quantity by their sum. Relative amounts of dispersion of the price changes in 1897 and in 1910, as shown in Table 52, when computed within the limits of the 1st and 9th deciles, are as follows:

$$1897: \frac{102 - 71}{102 + 71} = .18; \quad 1910: \frac{187 - 86}{187 + 86} = .37$$

The corresponding coefficients based upon the extremes are:

$$1897: \frac{128 - 56}{128 + 56} = .39; \quad 1910: \frac{363 - 48}{363 + 48} = .77$$

The effect of choosing the 1st and 9th deciles rather than the extremes is to reduce the relative dispersion by approximately one half.

Another method of showing dispersion by the method of limits, but of a somewhat different type from the selection of the extremes or a pair of deciles, is to take the ranges covered by *successive* tenths (deciles) in a series. This is done in an interesting way by Mitchell in the note on page 330.

The relation of the dispersion of one part of a statistical series compared with that of the whole may be determined by comparing the range of the middle fifty per cent of the cases with that of the total. For instance, the inventories as per cents of sales for the middle half of a group of retail clothing

stores fall within a range of one third of that covered by the entire group. That is, dispersion is much less for the part selected than for the entire series. By an extension of the

AVERAGE CONCENTRATION OF PRICE FLUCTUATIONS AROUND THE MEDIAN, 1891 TO 1913

[The fluctuations represent percentage changes from average prices in the preceding year.]

| AVERAGE RANGE COVERED BY THE— | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1st and 10th tenths of the price fluctuations | 2d and 9th tenths of the price fluctuations | 3d and 8th tenths of the price fluctuations | 4th and 7th tenths of the price fluctuations | 5th and 6th tenths of the price fluctuations | Successive tenths of the price fluctuations | Central two tenths of the price fluctuations | Central four tenths of the price fluctuations | Central six tenths of the price fluctuations | Central eight tenths of the price fluctuations | Whole number of the price fluctuations |
| | ...... | ...... | ...... | ...... | 1st tenth, 27.0 | | | | | |
| | | ...... | ...... | ...... | 2d tenth, 4.9 | | | | | |
| | | | ...... | ...... | 3d tenth, 2.6 | | | | | |
| | | | | ...... | 4th tenth, 2.2 | | | | | |
| 69.4 | 11.8 | 6.1 | 4.2 | 3.6 | 5th tenth, 1.8 | 3.6 | 7.8 | 13.9 | 25.7 | 95.1 |
| | | | | | 6th tenth, 1.8 | | | | | |
| | | | | ...... | 7th tenth, 2.0 | | | | | |
| | | | ...... | ...... | 8th tenth, 3.5 | | | | | |
| | | ...... | ...... | ...... | 9th tenth, 6.9 | | | | | |
| | ...... | ...... | ...... | ...... | 10th tenth, 42.4 | | | | | |

"The central division of the table shows that the average range covered by the fluctuations diminishes rapidly as we pass from the cases of greatest fall toward the cases of little change, and then increases still more rapidly as we go onward to the cases of greatest rise. The right-hand group of columns shows how the range increases if we start with the two middle tenths, take in the two tenths just outside them, then the two tenths outside the latter, and so on until we have included the whole body of fluctuations. The left-hand group of columns, on the other hand, combines in succession the two tenths on the outer boundaries, then the two tenths immediately inside them, and so on until we get back again to the two central tenths. Perhaps the most striking single result brought out by this table is that eight tenths of all the fluctuations are concentrated within a range (25.7 per cent) slightly narrower than that covered by the single tenth that represents the heaviest declines (27 per cent), and much narrower than that covered by the single tenth that represents the greatest advances (42.4 per cent)."

Mitchell, Wesley C., "Index Numbers of Wholesale Prices in the United States and Foreign Countries," *Bulletin of the United States Bureau of Labor Statistics*, No. 173, Washington, D. C., 1915, p. 17.

same method, the lower may be compared with the upper half; or any part with any other part. For many purposes such comparisons are illuminating.

When, for instance, the modal limits and the number of cases falling within them are given, and when the total range and the total number of cases are known, relative measures of the dispersion within the modal group as compared with that over the whole series may be computed. In the total section of Table 47, the modal group of 196 cases falls at 18-20 per cent. That is, it covers a range of 2 per cent. The range of the 1088 instances is 32 per cent. Accordingly, for the entire series there are on the average 34 cases, and for the modal group 98 cases for each one per cent of change. The dispersion over the entire series, therefore, is approximately three times as great as it is within the modal group.

### (2) *The Decile Method (Graphic) for Time Series*

The deciles may also be used to show graphically amounts of dispersion. Professor Mitchell has used them in two interesting ways: first, to show by years the dispersion of *relative* wholesale prices for 1890 to 1910; and second, to show by years the dispersion of the *change* in wholesale prices from 1891 to 1918.

In the first use,[1] the prices of 145 commodities in each year are computed as percentages of their prices in 1890 to 1899. That is, in each year there are 145 relative numbers or per cents. These are arranged in order of size each year and the nine deciles computed.[2] The deciles and the extremes are shown in Table 52. The amount of dispersion may be calculated arithmetically or shown graphically.

[1] Mitchell, Wesley C., *Business Cycles*, University of California Studies, Berkeley, 1913, p. 112.

[2] The formulæ for computing the 1st, 2nd, 7th deciles, respectively, are: $\frac{n+1}{10}$; $\frac{2(n+1)}{10}$; $\frac{7(n+1)}{10}$. In all cases $n$ refers to the number of items.

TABLE 52

TABLE SHOWING THE DECILES OF RELATIVE WHOLESALE PRICES IN THE UNITED STATES, BY YEARS — 1890-1910

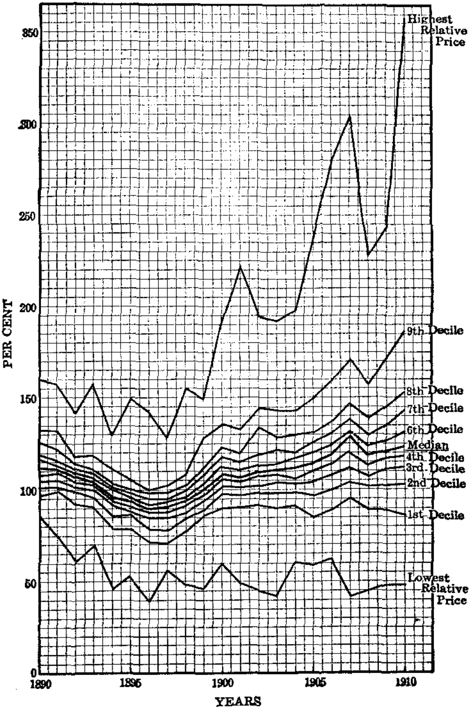
(Taken from Mitchell, W. C., *Business Cycles*, p. 112)

| Years | Lowest Relative Price | 1st Decile | 2nd Decile | 3rd Decile | 4th Decile | 5th Decile Median | 6th Decile | 7th Decile | 8th Decile | 9th Decile | Highest Relative Price |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1890 | 86 | 97 | 101 | 105 | 108 | 112 | 116 | 119 | 126 | 133 | 160 |
| 1891 | 74 | 99 | 101 | 105 | 109 | 111 | 113 | 116 | 122 | 132 | 158 |
| 1892 | 61 | 92 | 99 | 101 | 104 | 107 | 108 | 111 | 114 | 118 | 141 |
| 1893 | 70 | 90 | 96 | 100 | 102 | 104 | 106 | 109 | 111 | 119 | 158 |
| 1894 | 46 | 79 | 85 | 91 | 94 | 96 | 99 | 101 | 103 | 111 | 129 |
| 1895 | 53 | 79 | 86 | 88 | 91 | 94 | 95 | 98 | 100 | 105 | 149 |
| 1896 | 39 | 71 | 79 | 85 | 88 | 90 | 92 | 95 | 98 | 100 | 142 |
| 1897 | 56 | 71 | 78 | 85 | 88 | 91 | 93 | 95 | 98 | 102 | 128 |
| 1898 | 48 | 77 | 84 | 87 | 91 | 94 | 96 | 99 | 101 | 108 | 155 |
| 1899 | 46 | 86 | 89 | 94 | 97 | 100 | 103 | 108 | 112 | 129 | 149 |
| 1900 | 59 | 90 | 98 | 102 | 106 | 109 | 113 | 118 | 123 | 136 | 192 |
| 1901 | 49 | 90 | 97 | 101 | 104 | 107 | 111 | 115 | 120 | 133 | 222 |
| 1902 | 45 | 91 | 98 | 102 | 107 | 110 | 114 | 119 | 134 | 145 | 194 |
| 1903 | 43 | 90 | 98 | 104 | 108 | 111 | 114 | 121 | 129 | 143 | 192 |
| 1904 | 60 | 91 | 98 | 103 | 106 | 112 | 117 | 120 | 130 | 143 | 197 |
| 1905 | 59 | 85 | 97 | 104 | 110 | 114 | 120 | 126 | 131 | 149 | 238 |
| 1906 | 62 | 89 | 100 | 108 | 114 | 119 | 124 | 131 | 137 | 159 | 279 |
| 1907 | 42 | 95 | 104 | 112 | 121 | 129 | 132 | 139 | 147 | 171 | 304 |
| 1908 | 45 | 89 | 102 | 107 | 113 | 119 | 124 | 130 | 139 | 156 | 228 |
| 1909 | 48 | 89 | 102 | 111 | 117 | 121 | 127 | 135 | 146 | 172 | 243 |
| 1910 | 48 | 86 | 103 | 112 | 118 | 124 | 132 | 144 | 154 | 187 | 363 |

Concerning the amount of dispersion as shown by the table, Mitchell says: "In 1909, for example, one commodity had a relative price as low as 48, and another had a relative price as high as 243. Thus the arithmetic mean for that year, 121, represents relative prices which are scattered over a range of almost 200 points. But three-fifths of the 145 commodities had relative prices falling within a much narrower range—44 points, the difference between the second and eighth dec-

FIGURE 64

CURVES SHOWING, BY THE RANGE AND THE DECILE METHODS, THE DISPERSION OF THE FLUCTUATIONS IN RELATIVE WHOLESALE PRICES OF 145 COMMODITIES, 1890-1910



iles—and one-fifth fell within limits of ten points—the difference between the fourth and sixth deciles."[1]

A more effective method is to use a graphic device, such as Figure 64, on which are plotted each year the different deciles and the extremes. Dispersion each year is indicated by the distances on the ordinates within which the respective measures fall. As the different decile-lines converge, dispersion decreases; as they diverge, dispersion increases. A continuous and detailed picture is given of the spread or scatter.

The other graphic device used by Mitchell[2] in order to show dispersion by the decile method is reproduced in Figure 65. It is drawn on a logarithmic or ratio scale and

"shows for each year the whole range covered by the recorded changes from prices in the preceding year by vertical lines, which connect the points of greatest rise with the points of greatest fall. These lines differ considerably in length, which indicates that price changes cover a wider range in some years than in others. The heavy dots upon the vertical lines show the positions of the deciles. One-tenth of the commodities quoted in any given year rose above their prices of the year before by percentages scattered between the top of the line for that year and the highest of the dots. Another tenth fell in price by percentages scattered between the bottom of the line and the lowest of the dots. The fluctuations of the remaining eight-tenths of the commodities were concentrated within the much narrower range between the lowest and the highest dots. The dots grow closer together toward the central dot, which is the median. This concentration indicates, of course, that the number of commodities showing fluctuations of relatively slight extent was much larger than the number showing the wide fluctuations falling outside the highest and lowest deciles, or even between these deciles and the deciles next inside them.

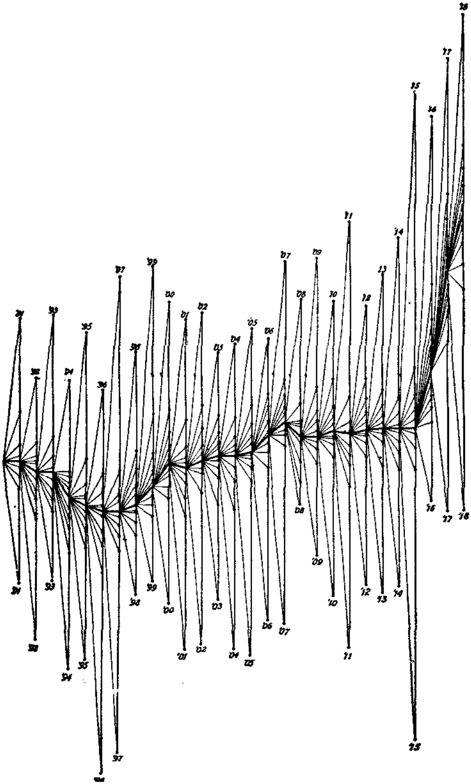
"The middle dots or medians in successive years are connected by a heavy black line, which represents the general upward or downward drift of the whole set of fluctuations. To make this drift

[1] *Op. cit.*, p. 109.

[2] Mitchell, Wesley C., "Index Numbers of Wholesale Prices in the United States and Foreign Countries," *Bulletin of the United States Bureau of Labor Statistics*, No. 284, Washington, D. C., 1921, p. 15, and chart facing it.

FIGURE 65

Conspectus of Yearly Changes in Prices, 1891-1918



clear the median of each year is taken as the starting point from which the upward or downward movements in the following year are measured. Hence the chart has no fixed base line. But in this respect it represents faithfully the figures from which it is made; since these figures are percentages of prices in the preceding year, a price fluctuation in any year establishes a new base for computing the percentage of change in the following year. The fact that prices in the preceding year are the units from which all the changes proceed is further emphasized by connecting the nine deciles, as well as the points of greatest rise and fall, with the median of the year before by light diagonal lines. The chart suggests a series of bursting bomb shells, the bombs being represented by the median dots of the years before and the scattering of their fragments by the lines which radiate to the deciles and the points of the greatest rise and fall."[1]

## 2. THE METHOD OF AVERAGING DIFFERENCES FROM A TYPE

The measures and coefficients of dispersion described above, while utilizing all or a part of the detail of statistical series, are not based upon any assumption as to the manner in which items are distributed about a norm or standard. No central term such as arithmetic mean, median, or mode is taken as a type from which divergence is summarized or averaged.

Those which are now to be described are quite different. Deviations or differences from a central type are totaled and averaged, and the amount of dispersion then expressed as a ratio to the standard selected. This general method, of which there are several modifications in current use, is based upon the assumptions (1) that statistical series tend to be distributed around their averages in a definite and regular manner, and, therefore, that an average is the appropriate standard from which to measure deviations (errors), and (2) that for such distributions the deviations so taken have certain mathemati-

[1] "Owing to the constant shifting of the base line, no fixed scale of relative prices can be shown on the margin of the chart. Because of its intricacy, the chart had to be reproduced on a larger scale than in the other cases, but of course that fact does not alter the slant of the lines, and this slant is the matter of importance."

cal properties which give the measures significance. Moreover, for ideal or probability distributions the different measures are related to each other by certain constants which it is desirable to utilize. What these are and the manner in which they are used will be developed as the different measures—absolute and relative—are described.[1]

### (1) *The Average Deviation*

The average deviation is exactly what its name implies—an average of the deviations. But what are deviations? Deviations from what? And what sort of an average is used to "average" them? For this measure, deviations are differences from a selected standard. This may be the arithmetic mean, the median, or the mode of the gross items. If a distribution is normal, these averages coincide, and it is a matter of indifference what name is applied to the norm taken. But most distributions are not of this type—they are non-symmetrical or skewed[2]—so that there is a difference between them. If deviations are taken from the arithmetic mean, their alegbraic sum equals zero, but since interest is in the amount of the deviations and not in their signs, all deviations are counted as positive.

But why choose the arithmetic mean rather than the median or the mode? One important reason for selecting the mean is because it is always a definite quantity while the median may be in doubt—there may be no actual quantity which divides a series into equal parts. Moreover, the mode may be ill-defined or there may be no mode at all. The deviations from the median, however, are smaller than those taken from any other quantity—that is, they are a minimum—and this is a desirable mathematical property of the deviations which it is desirable to use.[3]

[1] See Chapter XI, pp. 367-369.

[2] See Chapter XII.

[3] *By the use of an analogy*, Bowley has shown that the sum of the deviations is a minimum when calculated from the median. He says:

Accordingly, mathematical consistency seems to demand *that the median be used. But what is to be done if there is* no true median? This is often the case in discrete series. To measure the deviations from a median secured by interpolation may make the sum of the deviations greater or less than those secured by using the arithmetic mean. While ideally the median should be used, necessity often requires that the deviations be computed from the arithmetic mean.[1]

But the deviations, although taken from an average of some sort are themselves averaged. For this purpose, the arithmetic mean is customarily used. But why? Is not the median of the different deviations quite as suitable?[2] Why use an average at all? Why not express them in some form which will not average out the differences but which will develop the typical amounts? For the latter purpose the mode might be chosen, or even a frequency distribution employed. But the mode of the differences may be quite as uncertain in amount

(Note 3 continued)

". . . Suppose that it is required to run from a telephone exchange separate wires to every one of $n$ places in a straight line, where should the exchange be placed, so as to use the least total amount of wire? At the median position. For if you move from the median position to the right or to the left, you will find immediately that you are adding more wire than you are subtracting. Supposing there are 20 stations, and you have a position between the 10th and 11th; if you move to a position between the 11th and 12th, you have to increase your distance from 10 stations and diminish it from 9, in every case by the same length of the wire. The wires correspond to deviations; and the sum of lengths of the wires is the sum of the lengths of the deviations. Consideration of this illustration will show that the sum of the deviations is a minimum when they are measured from the median, but that the median is not quite determinate, for if there are an even number of stations, the sums of the deviations measured from all points between the two central stations are the same." Bowley, A. L., *Measurement of Groups and Series*, Layton, London, 1903, p. 30.

[1] In moderately asymmetrical distributions the difference in the aggregate in the two cases would be small; in those which are markedly skewed, it may be appreciable.

[2] The median of the deviations from the average, if they are all taken as positive, is equivalent *in a normal curve* of error to the "probable error." For explanation of this constant, see Chapter XI, pp. 370-374.

as that of the original items.[1] If precision is desired, the use of both a mode and a frequency distribution must be ruled out, and the customary method used. To take the arithmetic mean of the sum of the differences gives a definite quantity and reduces series with different frequencies to a comparable basis.

But like the average of the original items it is an average. It does not give the deviations in detail, but only records a type. When they are uniform and small, it does this satisfactorily. When they are large and different, it fails here as it does with the gross items. Moreover, it is impossible to determine from the average alone which condition obtains. To do so requires that they be arranged into frequency groups or that the method of cumulative- or moving-range be used. When this is necessary must be determined by the data and the purposes for which they are used.

In the following examples the method of computing the average deviation is fully illustrated.

### a. The Average Deviation in Historical Series

Table 53 gives the quantity of tin plates imported into the United States, 1906-1915, inclusive, in millions of pounds. By disregarding signs and combining the deviations the total is 502.8. The average is therefore 502.8 ÷ 10 = 50.28. That is, the average difference of the various amounts imported from the average imported is 50.28 million pounds. The average itself is 86.6 million pounds. In one year the average is exceeded by 67.4 million pounds, while in another year the average imported exceeds the amount brought in in that year by 79.6 million pounds. The excess of the first is 78 per cent, and the deficit of the second 92 per cent, of the average. The average difference is 58 per cent of the average imported.

These differences are illustrated in Table 54.

[1] Normally, the differences of the items in a series from an average are more alike than are the items themselves.

## TABLE 53

TABLE SHOWING THE QUANTITY OF TIN PLATES IMPORTED INTO THE UNITED STATES, 1906-1915, INCLUSIVE, IN MILLIONS OF POUNDS *

| YEARS | AMOUNT | FREQUENCIES | DEVIATIONS | | |
|---|---|---|---|---|---|
| | | | From average, 86.6 | | Total (signs ignored) |
| | | | − | + | |
| Total | 86.6 (av.) | 10 | 251.4 | 251.4 | 502.8 |
| 1906 | 121 | 1 | | 34.4 | 251.4 |
| 1907 | 143 | 1 | | 56.4 | |
| 1908 | 141 | 1 | | 54.4 | |
| 1909 | 117 | 1 | | 30.4 | |
| 1910 | 154 | 1 | | 67.4 | |
| 1911 | 95 | 1 | | 8.4 | |
| 1912 | 7 | 1 | 79.6 | | 251.4 |
| 1913 | 28 | 1 | 58.6 | | |
| 1914 | 49 | 1 | 37.6 | | |
| 1915 | 11 | 1 | 75.6 | | |

* *Statistical Abstract of the United States*, 1915, p. 498.

## TABLE 54

TABLE SHOWING IN CLASSIFIED FORM THE DIFFERENCES FROM THE AVERAGE IMPORTATIONS OF TIN PLATES INTO THE UNITED STATES

(Based on Table 53)

| DIFFERENCES FROM THE AVERAGE IMPORTATIONS (IN MILLION POUNDS) | YEARS IN WHICH THE CORRESPONDING DIFFERENCES WERE FOUND | | |
|---|---|---|---|
| | Total | − | + |
| Total 86.6 (average) | 10 | 4 | 6 |
| Less than 15.0 | 1 | — | 1 |
| 15 but less than 30.0 | — | — | — |
| 30 but less than 45.0 | 3 | 1 | 2 |
| 45 but less than 60.0 | 3 | 1 | 2 |
| 60 but less than 75.0 | 1 | — | 1 |
| 75 but less than 90.0 | 2 | 2 | — |

Summarizing this table, it is shown that the positive and the negative differences from the average range from 90 to below 15 million pounds, six of the frequencies, when the deviations are taken positively, being between 30 and 60 million. The median difference when interpolated for is 55.4.

The average deviation may also be computed from an assumed average. The following table using the above data illustrates the method:

TABLE 55

TABLE SHOWING THE METHOD OF COMPUTING THE AVERAGE DEVIATION WHEN AN ASSUMED AVERAGE IS USED

(Data same as in Table 53)

| YEAR | AMOUNT | FREQUENCIES | | DEVIATIONS FROM ASSUMED AVERAGE — 90 | | |
|---|---|---|---|---|---|---|
| | | | | − | + | Total (signs ignored) |
| *Total* | *866* | | *10* | *265* | *231* | *496* |
| 1906 | 121 | 1 | 6 | | 31 | 231 |
| 1907 | 143 | 1 | | | 53 | |
| 1908 | 141 | 1 | | | 51 | |
| 1909 | 117 | 1 | | | 27 | |
| 1910 | 154 | 1 | | | 64 | |
| 1911 | 95 | 1 | | | 5 | |
| 1912 | 7 | 1 | 4 | 83 | | 265 |
| 1913 | 28 | 1 | | 62 | | |
| 1914 | 49 | 1 | | 41 | | |
| 1915 | 11 | 1 | | 79 | | |

The total error in deviations is 34—the difference between 265 and 231. Had the deviations been computed from the true average the difference would have been zero. The average error is, therefore, 34 ÷ 10, or 3.4. The deviations for six of the frequencies are too small—they were computed from 90 in place of 86.6—and for four of them they are too large for the same reason. Therefore (6 × 3.4) — (4 × 3.4), or 6.8, must be added to the combined deviations, 496, to make up

for the error. This gives 502.8 as the correct sum of the deviations when taken positively. The average deviation is, therefore, 502.8 ÷ 10, or 50.28, as in the first method above.

There is no presumption of a normal or ideal arrangement in a time series. The average deviation, therefore, loses some of the significance associated with it in the treatment of natural phenomena. In the case of economic statistics it may be highly artificial. By its very nature the differences are important not only because of their size but also because of their distance from the center of gravity. In the example in Table 53, the deviation of 8.4 is as important in the divisor as is that of 79.6. Each constitutes one of the ten differences. Of course, the median and the mode are differently affected.[1]

### b. The Average Deviation in Frequency Series

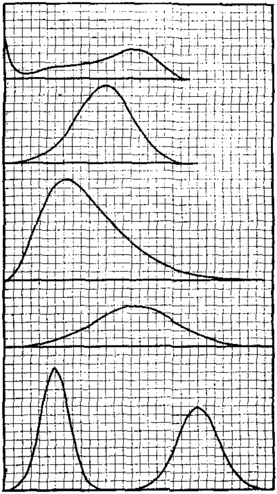
In the discussion of the average deviation for frequency series there is no necessity of restating the essential differences between those that are discrete and those that are continuous in type. What has already been said in this respect applies here. The present task is to comprehend its meaning and see its application to economic and business facts when they are grouped in frequency series.

Various types of frequency distributions are shown in Figure 66. Even on casual inspection, it is evident that it is futile to attempt to summarize them by a single expression such as an average. The averages may be similar, but the distributions about them widely different. It is the latter which are now being considered. Taking a somewhat different series, the application is seen in Table 56. Provided the signs are ignored, the differences amount to $50.65. The average difference is, therefore, $50.65 ÷ 37, or $1.37. That is, the *average difference from the arithmetic average is 32 per cent* of the average, and varies, when weighted according to its

[1] See what is said relative to this point in Chapter IX, *supra.*

FIGURE 66

Types of Frequency Distributions



mportance, from the smallest positive difference of $.54 to he largest negative difference of $11.07.

The manner in which the average deviation is computed for a grouped series is to assume for each group a uniform distribution of the frequencies, or what is the same thing, to assume that they are concentrated at the middle points, and pro-

TABLE 56

TABLE SHOWING THE METHOD OF COMPUTING THE AVERAGE DEVIATION IN A SIMPLE FREQUENCY DISTRIBUTION

| Amount | Frequencies | Deviations: From True Average, $4.23 − | + | Multiplied by the Frequencies − | + | Total (signs ignored) |
|---|---|---|---|---|---|---|
| Total | 37 | | | $25.33 * | $25.32 * | $50.65 |
| $2.00 | 4 | $2.23 | | 8.92 | | 8.92 |
| 4.00 | 3 | .23 | | .69 | | .69 |
| 3.00 | 9 | 1.23 | | 11.07 | | 11.07 |
| 6.00 | 5 | | $1.77 | | 8.85 | 8.85 |
| 3.00 | 2 | 1.23 | | 2.46 | | 2.46 |
| 8.00 | 3 | | 3.77 | | 11.31 | 11.31 |
| 5.00 | 6 | | .77 | | 4.62 | 4.62 |
| 3.50 | 3 | .73 | | 2.19 | | 2.19 |
| 4.50 | 2 | | .27 | | .54 | .54 |

* This negligible difference is due to taking the average as $4.23 rather than as $4.22 +.

ceed as in the case above. Table 57, using a different set of data, is illustrative.

The sum of the deviations is $610.60, and the average deviation $1.41. In this case, because of the concentration in the group $9.00 to $9.99, the average deviation is not much larger than the extent of this group, and is only 16 per cent of the average from which the deviations are computed. Moreover, the amount of dispersion in the frequency series in Table 57, relative to the average, is only one half as great as it is in the ungrouped series in Table 56. The clustering of the items at $9.00 to $9.99 shows that the average deviation is small, but it does not give it a numerical measure, nor does it localize it.

*TABLE 57*

TABLE SHOWING THE METHOD OF COMPUTING THE AVERAGE DEVIATION FROM A GROUP-FREQUENCY SERIES

| AMOUNTS | FREQUENCIES | DEVIATIONS | | | | |
|---|---|---|---|---|---|---|
| | | From the Average, $9.04 | | Product of Deviations and Frequencies | | Total Deviations (*signs ignored*) |
| | | − | + | − | + | |
| Total ....... | 434 | | | $305.48 * | $305.12 * | $610.60 |
| $5.00 to $5.99 | 15 | $3.54 | | 53.10 | | 53.10 |
| 6.00 to 6.99 | 40 | 2.54 | | 101.60 | | 101.60 |
| 7.00 to 7.99 | 66 | 1.54 | | 101.64 | | 101.64 |
| 8.00 to 8.99 | 91 | .54 | | 49.14 | | 49.14 |
| 9.00 to 9.99 | 113 | | $.46 | | 51.98 | 51.98 |
| 10.00 to 10.99 | 49 | | 1.46 | | 71.54 | 71.54 |
| 11.00 to 11.99 | 30 | | 2.46 | | 73.80 | 73.80 |
| 12.00 to 12.99 | 27 | | 3.46 | | 93.42 | 93.42 |
| 13.00 to 13.99 | 2 | | 4.46 | | 8.92 | 8.92 |
| 14.00 to 14.99 | 1 | | 5.46 | | 5.46 | 5.46 |

* This negligible difference is due to taking the average to be $9.04 rather than $9.039 +.

If the differences are calculated from an assumed average, it is necessary to make a correction for the difference between the guessed and the true average. The manner in which this is done in frequency series is shown in Table 58.

The total error in deviations is $200.00—the difference between $403.00 and $203.00. The average error is, therefore, $200.00 ÷ 434, or $.461. But the deviations of 212 of the frequencies are too large since they were computed from $9.50 instead of $9.04; and those of 222 are too small for the same reason. Therefore, the difference between 212 × $.461 and 222 × $.461 must be added to the total frequencies—$606.00—in order to get the correct total. $606.00 − (212 × $.461) +

(222 × \$.461) = \$610.60, and this divided by the number of instances, 434, equals \$1.41, the correct average deviation.

TABLE 58

TABLE SHOWING THE METHOD OF COMPUTING THE AVERAGE DEVIATION IN A GROUP-FREQUENCY SERIES WHEN AN ASSUMED AVERAGE IS USED

| Amounts | Frequencies | | Deviations: From Assumed Average, \$9.50 | | Product of Deviations and Frequencies | | Total Deviations (signs *ignored*) |
|---|---|---|---|---|---|---|---|
| | | | − | + | − | + | |
| Total ..... | | 434 | | | \$403.00 | \$203.00 | \$606.00 |
| \$5.00 to \$5.99 | 15 | 212 | \$4.00 | | 60.00 | | 60.00 |
| 6.00 to 6.99 | 40 | | 3.00 | | 120.00 | | 120.00 |
| 7.00 to 7.99 | 66 | | 2.00 | | 132.00 | | 132.00 |
| 8.00 to 8.99 | 91 | | 1.00 | | 91.00 | | 91.00 |
| 9.00 to 9.99 | 113 | 222 | | | | | |
| 10.00 to 10.99 | 49 | | | \$1.00 | | 49.00 | 49.00 |
| 11.00 to 11.99 | 30 | | | 2.00 | | 60.00 | 60.00 |
| 12.00 to 12.99 | 27 | | | 3.00 | | 81.00 | 81.00 |
| 13.00 to 13.99 | 2 | | | 4.00 | | 8.00 | 8.00 |
| 14.00 to 14.99 | 1 | | | 5.00 | | 5.00 | 5.00 |

The so-called "step-deviation" method, used in Chapter IX for computing the arithmetic mean, may be used in connection with the average deviation. Moreover, a consideration to be kept in mind when the method employed in Table 56 is used, may be explained. Suppose an average of \$10.50 is assumed and that the average deviation is calculated for the above series by the "step" method. Table 59 shows the result.

The total error in step-deviations is 634; the difference between 728 and 94. The average step-deviation error is,

therefore, 634 ÷ 434 or 1.46. The steps are all of \$1.00 width, so that the average step-deviation error, in terms of the unit of measurement, is \$1.00 × 1.46 or \$1.46. But the combined deviations, 822, are computed from \$10.50 instead of \$9.04, *the true average. Some of them are too small and some are* too large. Which are affected and how much? The deviations of the frequencies above \$8.50 are each too large by \$1.46 on the average. Those at \$10.50 and below are each too small by the same amount. Those at \$9.50, 113, are each too large by \$1.00 if \$10.50 is used. But, \$9.04 instead of \$9.50 is the average. Therefore, each of the 113 is too large by the difference between \$1.00 and \$.46, which is \$.54.[1] The total deviations properly corrected are 822 — (212 × \$1.46) + (109 × \$1.46) — (113 × \$.54) which equals \$610.6. The average deviation is, therefore, \$610.6 ÷ 434, or \$1.41.

This seems a roundabout method of reaching a simple result. It is, but only when the guessed average falls outside of the limits of the group which contains the true average. If it falls within this group, the method is simple and possesses merits for some uses.

So much for the method of computing the average deviation in both time and frequency series. Just a word of recapitulation. The average deviation is an *average*. It does *not* necessarily reflect the peculiarities of deviations any more

[1] The reason for an overlapping is shown by diagram below:

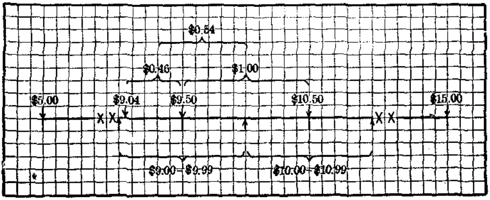


TABLE 59

TABLE SHOWING THE METHOD OF COMPUTING THE AVERAGE DEVIATION IN A GROUP-FREQUENCY SERIES FROM AN ASSUMED AVERAGE BY THE "STEP-DEVIATION" METHOD

| AMOUNTS | FREQUENCIES | | DEVIATIONS IN "STEPS" | | | | |
|---|---|---|---|---|---|---|---|
| | | | From Assumed Average, $10.50 | | Product of Deviations and Frequencies | | Total (signs ignored) |
| | | | − | + | − | + | |
| Total ...... | | 434 | | | 728 | 94 | 822 |
| $5.00 to $5.99 | 15 | 212 | 5 | | 75 | | 75 |
| 6.00 to 6.99 | 40 | | 4 | | 160 | | 160 |
| 7.00 to 7.99 | 66 | | 3 | | 198 | | 198 |
| 8.00 to 8.99 | 91 | | 2 | | 182 | | 182 |
| 9.00 to 9.99 | 113 | 113 | 1 | | 113 | | 113 |
| 10.00 to 10.99 | 49 | 109 | | | | | |
| 11.00 to 11.99 | 30 | | | 1 | | 30 | 30 |
| 12.00 to 12.99 | 27 | | | 2 | | 54 | 54 |
| 13.00 to 13.99 | 2 | | | 3 | | 6 | 6 |
| 14.00 to 14.99 | 1 | | | 4 | | 4 | 4 |

than the arithmetic mean does of data from which it is computed originally, except for the fact that the respective variations from the average deviation are usually not as large as are the variations of the original data from their average. If it is large it shows relative dispersion; if it is small it shows relative concentration. The exceptions are weighted in this case in the same way that they are in any arithmetic mean. If the median or modal deviations are used, then they exert less weight. If the cumulative-range method is used, they are thrown into prominence in detail.

Average deviations are reduced to a relative base by dividing them by the averages from which they are computed. By so doing they are reduced to a common denominator. Comparisons can then be made between dispersions in different series. This would be impossible by the use of measures of

dispersion alone for series in which the averages are unequal and for those expressed in different units. To divide the average deviation by the average produces a ratio or coefficient.

The relative dispersion in the frequency distribution used as an example is .156.[1] That is, it is the ratio secured by dividing $1.41—the average amount of dispersion—by $9.04 —the average from which dispersion of the items is measured.

*(2) The Standard Deviation*

The standard is a modification of the average deviation. It is computed (1) by taking the respective deviations from the arithmetic average, (2) by squaring them, thus getting rid of the minus signs, (3) by dividing the total by the number of frequencies, and (4) by extracting the square root of the quotient. In the formula, $n$ refers to the number of instances —frequencies; $d^2$, to the deviations squared: $\Sigma$ is the Greek capital letter S and means "the process of summation." In this case the amounts to be summated or totaled are the products of the frequencies and the squares. The standard deviation is usually indicated by small sigma, $\sigma$, or *S. D.* The formula by which it is calculated is $\sqrt{\frac{\Sigma\,(d^2)}{n}}$.

Squaring gives weight to extremes—those deviations far removed from the average. This is not fully compensated for in the subsequent root extraction. In frequency distributions which follow the normal law of error, or which are moderately asymmetrical, instances far removed from the average are relatively few, so that the products of the squares and the frequencies at these points are due more to the squaring than to the multiplication. Near the average, however, fre-

[1] On the graphic method of indicating absolute and relative dispersion, see Clark, Earle, "The Horizontal Zero in Frequency Diagrams," in *Quarterly Publications of the American Statistical Association*, June, 1917, pp. 662-669. This article is reprinted in the writer's *Readings and Problems in Statistical Methods*, Macmillan & Company, New York, 1920, pp. 385-394.

quencies are relatively numerous and the products affected by the concentration. In averaging the squares of the deviations, the frequencies, as such, exert equal weight, since the total is simply divided by the sum of the frequencies.

In time or historical series the case is somewhat different. There is no multiplication of deviations by frequencies, since each item appears but once. The squaring alone is effective. Of course, distance from the average is still important, but this is neither accentuated nor minimized by the distribution of frequencies. Just as the sum of the deviations is a minimum—that is, least—when calculated from the median, so the sum of the squares of the deviations is a minimum when calculated from the arithmetic mean. This follows from the principle that the nearest approach to the mathematically correct measure or observation in a series is the arithmetic mean, and that errors in observation are distributed about this center according to the rule of squares.[1]

For many economic and business purposes interest lies chiefly in the thing that is characteristic. Legislation is not generally enacted for the few, but rather for the many. Business policies are most frequently mapped out and changed in the light of that which seems to be characteristic. Sometimes, however, it is the exception which is suggestive, or which calls attention to the need for change. For instance, an exceptionally large sale—one far removed from the characteristic performance—may suggest possibilities in management and deserve to be emphasized both because of its stimulating effect on future performances on the part of salesmen, and because of its suggestive power to the management as to the need of reorganizing the selling force. Wide dispersion of employes' earnings in piece-work establishments may suggest to a keen business management the possibilities of redistributing his labor service according to capacity and proved ability. The losses resulting from a haphazard use of labor force, when

[1] See Yule, G. Udny, *Introduction to the Theory of Statistics*, Griffin, London, 1911, pp. 134-135.

measured in terms of discontent, turnover of labor, etc., may well make it advisable to assign more importance to the exception than that which would follow from its mere numerical significance. The inequalities of wealth distribution carry with them a significance far greater than that indicated by amounts alone.

So long as it is desired to give moderate weight to large differences, the average deviation may be used. When interest shifts to that which is exceptional, means of throwing it into light are needed. Of course, in statistics of economics and business there is generally not the same presumption of normal distribution as there is in statistics of natural phenomena. Interest in deviations from type in the two cases is of a different kind. Respecting the latter, deviations are important as showing non-conformity to an abstract standard; respecting the former, as means of calling attention, for instance, to useless waste, to unnecessary sources of industrial disorder, etc. Approach in the two cases may be different, but the means of measuring the concentration or dispersion is the same. To cite an average alone is frequently inadequate in economics, even for general purposes. But to use both an average and the standard deviation gives a rather definite idea of distribution about this figure. The latter serves more accurately to define the average. Moreover, average and standard deviations bear a more or less definite relation to each other in distributions which approach the normal law. As Yule says,

> "It is a useful empirical rule for the student to remember that for symmetrical or only moderately asymmetrical distributions, approaching the ideal forms . . ., the mean deviation is usually very nearly four-fifths of the standard deviation."[1]

Again, the standard deviation bears a more or less fixed relation to the total frequencies. Respecting this, Yule says:

[1] Yule, G. Udny, *Introduction to the Theory of Statistics*, Griffin, London, 1911, p. 146.

"It is a useful empirical rule to remember that a range of six times the standard deviation usually includes 99 per cent or more of all the observations in the case of distributions of the symmetrical or moderately asymmetrical type."[1]

How nearly this is true for the frequency distributions chosen for example is evident on inspection.

a. The Standard Deviation in Historical or Time Series

Using the time series of Table 53, the standard deviation is computed as follows, when the direct method is used:

TABLE 60

TABLE SHOWING THE METHOD OF COMPUTING THE STANDARD DEVIATION FOR HISTORICAL SERIES USING THE DIRECT METHOD

(Data same as in Table 53)

| YEARS | AMOUNT | FREQUENCIES | DEVIATIONS | | | |
|---|---|---|---|---|---|---|
| | | | From Average, 86.6 | | Squared | Squared, Multiplied by Frequencies |
| | | | − | + | | |
| Total | 86.6 (av.) | 10 | | | | 29,760.40 |
| 1906 | 121 | 1 | | 34.4 | 1,183.36 | 1,183.36 |
| 1907 | 143 | 1 | | 56.4 | 3,180.96 | 3,180.96 |
| 1908 | 141 | 1 | | 54.4 | 2,959.36 | 2,959.36 |
| 1909 | 117 | 1 | | 30.4 | 924.16 | 924.16 |
| 1910 | 154 | 1 | | 67.4 | 4,542.76 | 4,542.76 |
| 1911 | 95 | 1 | | 8.4 | 70.56 | 70.56 |
| 1912 | 7 | 1 | 79.6 | | 6,336.16 | 6,336.16 |
| 1913 | 28 | 1 | 58.6 | | 3,433.96 | 3,433.96 |
| 1914 | 49 | 1 | 37.6 | | 1,413.76 | 1,413.76 |
| 1915 | 11 | 1 | 75.6 | | 5,715.36 | 5,715.36 |

The deviations squared and totaled amount to 29,760.40. The standard deviation is, therefore, $\sqrt{\frac{29,760.40}{10}}$ or $\sqrt{2,976.04}$ or 54.5. The average deviation, 50.28, is 92.3 per cent of this amount.

[1] *Ibid.*, p. 140.

In Table 61, the deviations are taken from the assumed average, 90.0, instead of the true average, 86.6. The average error in the deviations is, therefore, 3.4. This must be squared, multiplied by the number of frequencies, and then subtracted from 29,876 in order to get the correct deviations squared. The square of 3.4 is 11.56, and when multiplied by 10—the number of frequencies—is 115.6. The difference between this amount and 29,876 is 29,760.4. The square root of this amount, 54.5, is the standard deviation. The problem is somewhat simplified by taking the deviations from an assumed average because the items to be squared are whole numbers. Of course, in actual work it is unnecessary to multiply the deviations by the frequencies since they are all unity. It was done here in order that all the steps might be followed.

TABLE 61

TABLE SHOWING THE METHOD OF COMPUTING THE STANDARD DEVIATION FOR HISTORICAL SERIES USING THE DIRECT METHOD BUT AN ASSUMED AVERAGE

(Data same as in Table 53)

| YEARS | AMOUNT | FREQUENCIES | DEVIATIONS | | | |
|---|---|---|---|---|---|---|
| | | | From Assumed Av., 90.0 | | Squared | Squared, *Multiplied* by Frequencies |
| | | | − | + | | |
| Total | 86.6 (av.) | 10 | | | | 29,876 |
| 1906 | 121 | 1 | | 31 | 961 | 961 |
| 1907 | 143 | 1 | | 53 | 2,809 | 2,809 |
| 1908 | 141 | 1 | | 51 | 2,601 | 2,601 |
| 1909 | 117 | 1 | | 27 | 729 | 729 |
| 1910 | 154 | 1 | | 64 | 4,096 | 4,096 |
| 1911 | 95 | 1 | | 5 | 25 | 25 |
| 1912 | 7 | 1 | 83 | | 6,889 | 6,889 |
| 1913 | 28 | 1 | 62 | | 3,844 | 3,844 |
| 1914 | 49 | 1 | 41 | | 1,681 | 1,681 |
| 1915 | 11 | 1 | 79 | | 6,241 | 6,241 |

### b. The Standard Deviation in Frequency Series

The method of calculating the standard deviation is the same for frequency as for time series, but it may be helpful to carry through an example when the direct and the indirect methods are employed. Taking the data in Table 58, and assuming the average to be $9.50—the true average being $9.04—the short-cut method is as shown in Table 62.

The sum of the squares of the deviations from the guessed or assumed average is $1,424.00. But the average error is $.461. The square of $.461 is $.212. This amount multiplied by the number of frequencies—434—gives $92+, and this amount, when subtracted from $1424, gives $1332, as the correct deviations squared. But since it is the average of the squared deviations that is desired, it is necessary to divide

TABLE 62

TABLE SHOWING THE METHOD OF COMPUTING THE STANDARD DEVIATION FOR FREQUENCY SERIES BY USING THE SHORT-CUT METHOD AND AN ASSUMED AVERAGE

(Data same as in Table 58)

| AMOUNTS | FREQUENCIES | DEVIATIONS | | | |
|---|---|---|---|---|---|
| | | From Assumed Av., $9.50 | | Squared | Squared, Multiplied by Frequencies |
| | | − | + | | |
| Total....... | 434 | | | | $1,424.00 |
| $5.00 to $5.99 | 15 | $4.00 | | $16.00 | 240.00 |
| 6.00 to 6.99 | 40 | 3.00 | | 9.00 | 360.00 |
| 7.00 to 7.99 | 66 | 2.00 | | 4.00 | 264.00 |
| 8.00 to 8.99 | 91 | 1.00 | | 1.00 | 91.00 |
| 9.00 to 9.99 | 113 | | | | |
| 10.00 to 10.99 | 49 | | $1.00 | 1.00 | 49.00 |
| 11.00 to 11.99 | 30 | | 2.00 | 4.00 | 120.00 |
| 12.00 to 12.99 | 27 | | 3.00 | 9.00 | 243.00 |
| 13.00 to 13.99 | 2 | | 4.00 | 16.00 | 32.00 |
| 14.00 to 14.99 | 1 | | 5.00 | 25.00 | 25.00 |

this number by 434. The result is \$3.07. The square root of \$3.07, \$1.75, is the standard deviation. The average deviation—\$1.41—is 81 per cent of this amount.

The standard deviation of a series is somewhat larger than its average deviation. If the distribution is normal in the probability sense, the two measures of variability stand in the following relation:

$$\sigma \text{ or } S.\ D. = 1.2533\ A.\ D., \text{ or conversely,}$$

$$A.\ D. = 0.7979\ \sigma \text{ or } S.\ D.$$

Applying this formula to the example used as an illustration, the relation between the average and the standard deviations is as *1 : 1.2413*, or conversely, *0.8056 : 1*. That is, the distribution approaches very nearly the normal or probability type.

If the same distribution and a guessed average are used and the deviations are taken in terms of "steps," the method is the same, except that it is necessary to convert the steps into terms of the unit employed by multiplying by the size of the group. In this case the step is \$1.00. If the widths of groups had been \$.50, for instance, the conversion would have been made by multiplying the number of steps by one half dollar.

If deviations from the actual average, as they appear in Table 57, are used, the process is the same but somewhat more laborious to carry through since the deviations to be squared are not whole numbers. Of course, in such a case it is unnecessary to make a correction for errors in the deviations. They are correct by assumption.

In order to convert the standard deviation into a *coefficient* —that is, to relieve the data of the particular unit in which they are expressed, and to make comparisons possible between two series in which absolute units are different—it is only necessary to divide by the arithmetic mean—the figure from which the deviations are computed. The *coefficient* of dispersion for this series based on *S. D.*, is $\frac{\$1.75}{\$9.04}$, or .194.

### (3) *The Quartile Measure*

The quartile measure of dispersion applies to that portion of a distribution contained between the first and third quartiles. The extremes below the first and beyond the third quartiles are ignored. It serves to characterize that portion which lies nearest the average or type. This measure, like the average and standard deviations, is an average. It is not, however, calculated from the differences of the items from the arithmetic mean. By taking one half of the range contained in the middle half of a distribution, the measure shows the average deviation of the quartiles from the median. The formula is $\frac{Q3 - Q1}{2}$, where $Q3$ and $Q1$ stand for the third and first quartiles, respectively. The third quartile lies above the median; the first one below it. One half of all the frequencies lies between them. This measure is known as the semi inter-quartile range or quartile deviation and is frequently indicated by $Q$. In distributions which are symmetrical, the amounts secured by the use of the formula when added to the lower or subtracted from the upper quartile give the median. In those which are asymmetrical, such an amount may be greater or less than the median, depending upon the type of asymmetry. Because this measure, although based upon the method of limits, is used in connection with the median—a type amount—it is discussed here rather than in the section of the chapter devoted to the *Method of Limits.*

In symmetrical or moderately asymmetrical distributions the relation between the quartile and the standard deviation measures of dispersion is fairly constant and predictable. The first is generally about two thirds of the second, and nine times the first usually contains about 99 per cent of the range covered by the entire distribution.[1] How nearly this relation obtains in the distribution chosen as an illustration is shown by the following compilations: In Table 43, the median, by

[1] Yule, *op. cit.*, p. 148.

interpolation, is fixed at \$9.049. The first and third quartile positions, by the formula $\frac{n+1}{4}$, and $\frac{3\ (n+1)}{4}$, respectively, are the 108¾th and 326¼th men. The wages of these hypothetical individuals, when interpolated for, are \$7.81 and \$10.03, respectively. The quartile range is, therefore, \$10.03 — \$7.81, or \$2.22. The average range is $\frac{\$2.22}{2}$, or \$1.11.[1] For the same series the average deviation is \$1.41, and the standard deviation \$1.75. The semi inter-quartile range, therefore, is equal to 79 per cent of the former and 63 per cent of the latter. The extreme range of \$10.00—the difference between \$5.00 and \$15.00—is almost exactly nine times the quartile measure, \$1.11.

Like other measures of dispersion the semi inter-quartile range may be reduced to a relative basis, or made a coefficient, by dividing through by a common denominator. In this case, the appropriate divisor is the sum of the quartiles. The fraction —$\frac{Q3 - Q1}{Q3 + Q1}$— increases with the distance between the quartiles but always lies between 0 and 1. Size, therefore, is a test of relative dispersion. In the above example the coefficient is $\frac{\$10.03 - \$7.81}{\$10.03 + \$7.81}$, or .124. That is, the dispersion is relatively small. It is 79 per cent of the coefficient based on the average deviation and 64 per cent of the coefficient based on the standard deviation.

For many purposes a study of the semi inter-quartile range is sufficient. This may result from the nature of a distribution or from lack of interest in the extreme cases. However, to cite only this measure may prejudice a case for all purposes except those which are under discussion. In order to

[1] For discrete series, interpolation in units less than those in which data are measured is illogical and aims at too great accuracy. For most purposes the quartiles would be given with sufficient accuracy as \$7.80 and \$10.00.

guard against misunderstanding and to give expression to all of the peculiarities of a distribution, it is generally better to determine the average, the standard, and the quartile deviations. A comparison of these gives an accurate picture of a distribution.

## IV. Summary

Measures and coefficients of dispersion serve more accurately to describe statistical series than is possible by the use of averages alone. They are more refined statistical summaries, *the amounts with which they have to do being the* differences of the items one from another, or from a standard which is considered typical or representative. When using them in historical series, nothing can be implied about the type of distribution. In frequency series, on the other hand, the selection of a type from which to measure the deviations suggests some natural or normal order of distribution. Moreover, the relations between the constants for normal curves establish a standard by which those found in individual cases may be judged or appraised. But what does the use of these constants imply? What are meant by such expressions as the "normal law of error curve," "a normal distribution"? Briefly to answer these questions is the subject of the following chapter.

### References

Bowley, A. L., *Elements of Statistics*, 4th Edition, King, London, 1920, Chapter VI, pp. 110-124.

Clark, Earle, "The Horizontal Zero in Frequency Diagrams," in *Quarterly Publications of the American Statistical Association*, June, 1917, pp. 662-669.

Davenport, Eugene, *Principles of Breeding*, Ginn and Co., New York, 1907, Chapter XII, pp. 419-452.

Elderton, W. P., and E. M., *Primer of Statistics*, Black, London, 1910, Chapter IV, pp. 40-55.

Jones, D. Caradog, *A First Course in Statistics*, Bell, London, 1921, Chapter VI.

Kelley, T. L., *Statistical Method,* Macmillan & Co., New York, 1923, Chapter IV.

King, W. I., *Elements of Statistical Method,* Macmillan & Co., New York, 1912, Chapters XIII and XIV, pp. 141-167.

Mills, Frederick C., *Statistical Methods Applied to Economics and Business,* Holt, New York, 1924, Chapter V, pp. 147-168.

Mitchell, W. C., "Methods of Presenting Statistics of Wages," *Quarterly Publications of American Statistical Association,* Vol. IX, pp. 325-343.

Mitchell, W. C., "Index Numbers of Wholesale Prices in the United States and Foreign Countries," *Bulletin* 284, *United States Bureau of Labor Statistics,* Washington, D. C., 1921, pp. 11-23.

Pearl, Raymond, *Introduction to Medical Biometry and Statistics,* Saunders, Philadelphia, 1923, Chapter XIII.

Rietz, H. L. (Editor) *Handbook of Mathematical Statistics,* Houghton Mifflin, Boston, 1924, Chapter II.

Rugg, H. O., *Statistical Methods Applied to Education,* Houghton Mifflin, Boston, 1917, Chapter VI, pp. 149-178.

Thorndike, E. L., *Theory of Mental and Social Measurements,* Columbia University, New York, 1916, Chapters III and IV, pp. 28-42 and 42-63, respectively.

Yule, G. U., *An Introduction to the Theory of Statistics,* Griffin, London, 1911, Chapter VIII, pp. 133-156, Sections 1, 2, 5 to 11, 13 to 30.

Žižek, Franz, *Statistical Averages* (translated by W. M. Persons), Holt, New York, 1913, Part III, Chapters I and II, pp. 251-255 and 256-270, respectively.

# CHAPTER XI

## THE THEORY OF PROBABILITY AND SOME PROPERTIES OF THE NORMAL LAW OF ERROR DISTRIBUTION[1]

### I. Outline of the Theory of Probability

In the measurements of natural and physical phenomena one is struck both by the similarities and the differences in different members of a class, or in repeated measurements of the same class. While results vary, they fall within clearly defined limits. In the absence of bias on the part of the one making the measurements, of changes in the unit, of the accuracy at which he aims, of the nature of the thing measured and of the unit in which the results are stated, there tends to be a common or typical measurement from which others deviate above and below in a more or less regular and systematic manner.

To illustrate: If the heights of a large number of a homogeneous class of men—say soldiers[2]—are measured to the

[1] A discussion of only the simplest phases of these subjects is suitable to an introductory text on statistical methods. The theory of probability belongs in the realm of mathematics as do also the more serious discussions of the properties of the normal law of error distribution. Both subjects are fully treated in the following among other books: Fisher, Arne, *The Theory of Probability*, Macmillan & Co., New York, 1922; Keynes, J. M., *A Treatise on Probability*, Macmillan & Company, London, 1921; less complete discussions are found in Pearl, Raymond, *Introduction to Medical Biometry and Statistics*, Saunders, Philadelphia, 1923, Chap. XI; Jones, D. C., A First Course in Statistics, Bell & Sons, London, 1921, Chaps. XII, XIII, and XVIII; Jevons, W. Stanley, *Principles of Science*, Macmillan & Company, London, 2nd Edition, 1920, Chap. X.

[2] See Yule, G. U., *An Introduction to the Theory of Statistics*, Griffin, London, 1911, Chap. VI for frequency graphs of measurement of heights of 1078 "English sons": 1,000 Cambridge Students; weight of 7,749 adult males in the British Isles, etc.

nearest quarter of an inch, differences will be found. Some men who may be termed "tall," by any reasonable standard, will be encountered. Similarly, some who are "short" will be found. The measurements, however, will cluster at or around a certain height which may be called modal. If, on the other hand, a non-homogeneous group of people—such for instance as that found at a Fair on a given day—were measured in the same way, the distribution of the results would be different. Those who are "short" for one class would be "tall" for another. Moreover, there is no necessary basis for expecting the heights definitely to cluster at a certain typical or modal measurement and shade off gradually above and below. A distribution of such an aggregate would probably have two modes. The same thing would be true of the sales of salesmen in 5 and 10 cent stores and of those in the furniture sections of department stores. Why? Because iu this and the foregoing example the phenomena are non-homogeneous.

Moreover, if the measurements of a homogeneous soldier group were made by several individuals with different standards of accuracy, affected by personal bias, or with non-uniform units of measurement, the results would not cluster about a type, and shade off systematically above and below. Why? Because the conditions of measurement are not uniformly applied.

To take another illustration. If the weights of a sufficiently homogeneous "population" of hogs at the Chicago stockyards were taken at a given time, the measurements being free from bias affecting the unit of measurement, the standard of accuracy and the sensitiveness of the scales, the weights would cluster about a norm or typical amount. If, on the other hand, they were taken over a period of time, during which methods of breeding, fattening, shipping, etc., made the receipts non-homogeneous, then no such type of distribution could be expected. Why? Because time has introduced an element of non-homogeneity or bias.

If, rather than measuring different members of a class a

number of times, a single example is subjected to many measurements, then, in the absence of bias affecting the purpose, intent, and prejudice of the one making the measurement, or the unit which is employed for this purpose, the normal type of distribution or a close approximation to it would result. Since by hypothesis accuracy is aimed at and non-homogeneous conditions—bias in every form—are removed, a typical or characteristic result would be secured. From this, however, there would be both negative and positive deviations since absolute uniformity is not to be expected. But these would be fewer in number than those which are termed characteristic or most common.

Similar illustrations drawn from other fields might be cited at length, but they would not add materially to the point which is being developed.

Let us approach the subject from a different angle. If a coin is tossed it may fall either heads or tails. It must fall *either* heads or tails; it cannot fall both in a single trial. If there is no reason why it should fall one way in preference to the other, it is said that the chances are even that the results will be heads or tails. If it is unevenly balanced, the head side being more heavily weighted, it will probably more frequently fall tails. That is, bias—in the coin itself—controls the results. If it is evenly balanced, but cleverly thrown, heads may markedly exceed tails. Bias in this case is personal. Chance—the name for that multitude of influences by which a given event is determined but all of which are supposed to operate without hindrance or bias—is interfered with.

Again, if cards are not evenly cut, equally smooth and of the same color and size, any one may be selected at will from a pack. If they are uniform in every particular, the chance of selecting a certain one is no greater than that of selecting any other one. If there are 52 in a pack, the chance of selecting the king of spades is 1 out of 52. *Some* card is selected, and since there are 52 possibilities, the chance of getting any one is the same as that of getting any other. Again, since

there are 13 diamonds, the chance of selecting some one diamond is 13 out of 52, or 1 in 4. But a diamond might not be selected if four trials were made, after each of which the card taken out is returned and the pack thoroughly shuffled. One might not be secured even if eight trials under the same conditions were made. But such a result would be very unlikely. *On the average*, with repeated drawings, one diamond out of each of four trials would tend to be selected. That is, the "probability" is ¼ that such a result would be secured. In the *long run* this would *tend* to be true.

Let us return to the illustrations of tossing coins. Suppose one coin is tossed a number of times in succession (an analogous case to measuring the same phenomenon a number of times). What is the probability of getting a certain number of heads and a certain number of tails?

In one toss, we may get either heads or tails. The chances, we say, are equal that one or the other result will be secured. Let the possible results be indicated as follows—H meaning heads, and T, tails:

H, T

In tossing the coin twice there can be four possible results. We can get

H H, H T, T H, T T

That is, a head in the second may follow a head in the first; a tail in the second, a head in the first; a head in the second, a tail in the first; and a tail in the second, a tail in the first.

In three tossings, there are eight possible results, because to the four events previously possible, the H and T of the third coin may be combined. These may be set down—using the same methods as above—as follows:

H H H, H H T, H T H, H T T, T H H, T H T, T T H, T T T

Similarly, with four tossings. In this case there are 16 possible events.

| | | | | |
|---|---|---|---|---|
| HHHH | HHHT | HHTT | HTTT | TTTT |
| | HHTH | THHT | THTT | |
| | HTHH | TTHH | TTHT | |
| | THHH | HTTH | TTTH | |
| | | HTHT | | |
| | | THTH | | |

If in place of writing the H's and T's separately we write as an exponent the number of times H and T appear in each combination, we get

$$H^4 + 4\,H^3T + 6\,H^2T^2 + 4\,H\,T^3 + T^4, \text{ or}$$
$$1 + 4 + 6 + 4 + 1$$

This is the number of ways in which the five combinations *can appear by tossing one coin four times. If four coins* were thrown once (this is an analogous case to measuring each of four things once) the result would be as follows—the different coins being designated as (a), (b), (c), (d):

| (a)(b)(c)(d) | (a)(b)(c)(d) | (a)(b)(c)(d) | (a)(b)(c)(d) | (a)(b)(c)(d) |
|---|---|---|---|---|
| HHHH | HHHT | HHTT | HTTT | TTTT |
| | HHTH | THHT | THTT | |
| | HTHH | TTHH | TTHT | |
| | THHH | HTTH | TTTH | |
| | | HTHT | | |
| | | THTH | | |

If each of these combinations is given an index notation, the result is the same as that secured when 1 coin is tossed 4 times, viz:

$$H^4 + 4\,H^3T + 6\,H^2T^2 + 4\,H\,T^3 + T^4, \text{ or}$$
$$1 + 4 + 6 + 4 + 1$$

Now this expression gives the same result as is obtained by raising the binomial $(H + T)$ to the fourth power. If it were raised to the fifth power the corresponding number of cases would be 32, made up of $1 + 5 + 10 + 10 + 5 + 1$, each of the H's and T's in the preceding example being combined with another H and T, thus producing twice as

many possible results. If it were raised to the 8th power, the number of possible events would be $1 + 8 + 28 + 56 + 70 + 56 + 28 + 8 + 1$.

From the "arithmetical triangle"[1] the number of times each combination may appear may be read off directly.[2]

It will be noted that each line of the "triangle" produces a series which regularly increases and then decreases, reaching a maximum at the center and shading off above and below.[3] *This is the probability distribution approached in the measurement of natural and physical phenomena.*

An illustration from Jevons at this place is of interest.

"Suppose, for the sake of argument, that all persons were naturally of the equal stature of five feet, but enjoyed during youth seven independent chances of growing one inch in addition. Of these seven chances, one, two, three, or more, may happen favorably to any indi-

[1] The Arithmetical Triangle

| Line | First Column | Second Column | Third Column | Fourth Column | Fifth Column | Sixth Column | Seventh Column | Eighth Column | Ninth Column | Tenth Column | Eleventh Column |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | | | | | |
| 2 | 1 | 1 | | | | | | | | | |
| 3 | 1 | 2 | 1 | | | | | | | | |
| 4 | 1 | 3 | 3 | 1 | | | | | | | |
| 5 | 1 | 4 | 6 | 4 | 1 | | | | | | |
| 6 | 1 | 5 | 10 | 10 | 5 | 1 | | | | | |
| 7 | 1 | 6 | 15 | 20 | 15 | 6 | 1 | | | | |
| 8 | 1 | 7 | 21 | 35 | 35 | 21 | 7 | 1 | | | |
| 9 | 1 | 8 | 28 | 56 | 70 | 56 | 28 | 8 | 1 | | |
| 10 | 1 | 9 | 36 | 84 | 126 | 126 | 84 | 36 | 9 | 1 | |
| 11 | 1 | 10 | 45 | 120 | 210 | 252 | 210 | 120 | 45 | 10 | 1 |

[2] See Jevons, W. Stanley, *Principles of Science* (2nd Edition), Macmillan & Company, London (Reprint 1920). "In general language, if I wish to know in how many ways $m$ things can be selected in combina*tions out of $n$ things, I must look in the $n + 1^{th}$ line, and take the* $m + 1^{th}$ number, as the answer. In how many ways, for instance, can a sub-committee of five be chosen out of a committee of nine. The answer is 126, and is the sixth number in the tenth line; it will be *found equal to* $\frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}$." *Ibid.*, p. 187.

[3] In alternate series above that for the 3rd power the two middle items are the same. See Jevons, *op. cit.*, pp. 185-186, for certain other properties of the "Arithmetical Triangle."

vidual; but, as it does not matter what the chances are, so that the inch is gained, the question really turns upon the number of combinations of 0, 1, 2, 3, etc., things out of seven. Hence the eighth line of the triangle gives us a complete answer to the question. . . . There are altogether 128 ways in which seven causes can be present or absent. Now, twenty-one of these combinations give an addition of two inches, so that the probability of a person under the circumstances being five feet two inches is $\frac{21}{128}$. The probability of five feet three inches is $\frac{35}{128}$; of five feet one inch $\frac{7}{128}$; of five feet $\frac{1}{128}$, and so on. Thus the eighth line of the Arithmetical Triangle gives all the probabilities arising out of the combinations of seven causes."[1]

The theoretical number of times different combinations of heads and tails would be secured if ten coins were tossed is shown in Table 63.

TABLE 63

THE THEORETICAL DISTRIBUTION SECURED BY TOSSING TEN COINS

(The 11th line of the Arithmetical Triangle)

| *Character of Throw* | *Theoretical Numbers* * |
|---|---|
| 10 Heads 0 Tails | 1 |
| 9 " 1 " | 10 |
| 8 " 2 " | 45 |
| 7 " 3 " | 120 |
| 6 " 4 " | 210 |
| 5 " 5 " | 252 |
| 4 " 6 " | 210 |
| 3 " 7 " | 120 |
| 2 " 8 " | 45 |
| 1 " 9 " | 10 |
| 0 " 10 " | 1 |
| Total.............. | 1,024 |

* See the 11th line of the Arithmetical Triangle.

Now upon the comparative number of combinations, as shown in the arithmetical triangle, as Jevons says, is founded

[1] *Op. cit.*, pp. 188, 202.

the theory of error [1] to which appeal is made in quantitative investigations.[2] The greater the number of times a group of coins is tossed, or the greater the number of coins which are tossed once, the nearer does the distribution of the results actually secured tend to agree with the theoretical distribution as given by the expansion of the binomial (H + T). Similarly, with perfect random selection the greater the number of natural phenomena of a homogeneous type which is measured once, as well as the greater the number of times a single phenomenon is measured, the nearer do the results secured agree with those which would characterize the entire "population." Upon the assumption that chance in the first instance and perfect random selection in the second produce the distribution in the arithmetical triangle, a theory of error is built up.

## II. Properties of the Normal Law of Error Distribution

If the theoretical results of tossing ten coins, as shown in Table 63, are plotted with the frequencies measured on the $Y$ axis, and the nature of combination on the $X$ axis we get Figure 67.

The ends of the ordinates are joined together by a smoothed line which would be approached if the exponent of the power of the binomial were increased—say to 999.[3] Figure 67 illustrates the so-called normal probability curve or normal law of error distribution to which reference has been made at various times. The shape of the curve is different for different exponents of the expansion of the bionomial. The lower the exponent, the more "peaked" the curve; the higher, the flatter

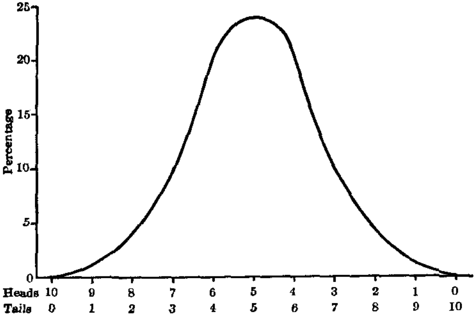
[1] The term "error" is used in the sense that if a number of observations are taken, the deviation or difference of any one of them from their mean is an "error."

[2] *Op. cit.*, pp. 153-189.

[3] For a figure in which the separate ordinates give essentially a smooth curve, see Slichter, Charles S., *Elementary Mathematical Analysis*, 2nd Edition, McGraw-Hill Book Co., New York, 1918, p. 212.

FIGURE 67

GRAPHICAL REPRESENTATION OF THE THEORETICAL DISTRIBUTION SECURED BY TOSSING TEN COINS




it appears. In all cases, however, the curves are alike in that they are symmetrical about a maximum—excesses and defects being equal—and shade off in a systematic and regular manner. Accordingly, such figures have certain mathematical properties of which the following are the most important:

1. The curve is uni-modal.
2. All of the instances are included beneath the curve and above the $X$ axis.
3. Half of the instances are included on either side of the mean.
4. The arithmetic mean, median, and mode coincide—they are identical.
5. The standard deviation, *S.D.*, cuts the curve at the points of inflection. Within a distance of one standard deviation, *S.D.*, above and below the mean 68 per cent of the instances fall.

6. Within a range of 2/3, or more exactly .6745, of the standard deviation, *S.D.*, when measured plus and minus from the mean, one half of all of the instances occur. This is the "probable error"—an expression which means that the chances are even that a measure (error or deviation from the mean) will fall within this interval.
7. The average deviation, *A.D.*, is four fifths—or more exactly .7979—of the standard deviation, *S.D.*
8. The semi inter-quartile range, $\frac{Q3 - Q1}{2}$, is equal to the probable error, *P.E.*—that is, a distance above and below the mean within which one half of the instances fall.
9. The semi inter-quartile range, $\frac{Q3 - Q1}{2}$, when added to the lower quartile or when subtracted from the upper quartile is equivalent to the mean, median, and the mode, and equal to 2/3 of the standard deviation, *S.D.*
10. The probable error, *P.E.*, is .845 of the average deviation, *A.D.*

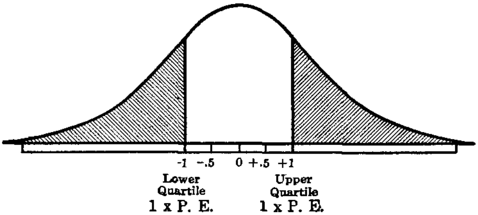
For the series showing the theoretical result of throwing ten coins, the arithmetic mean, median, and mode are 5; the standard deviation, *S.D.* or $\sigma$, is 1.58 and the probable error, 1.07. That is, it is an even chance that an item selected purely at random will fall within 5.00 ± 1.07 or between 3.93 and 6.07. The width of the shaded portion on Figure 68 shows the limits defined by the probable error, its area being one half of the total area under the curve.

Not only may the gross items of chance series or the measurement of phenomena taken at random be plotted in the form of a probability curve, but the different means (averages) of a number of such chance series or measurements may also be indicated in this manner. The means of different

measurements like the measurements themselves will vary.[1] If these were plotted as a frequency distribution, the form of graph would approach the normal type. Such a series would have a mean, a standard deviation, a probable error, etc., between which the relations may be expressed by a series of constants, in the same manner as for the gross items. For instance, the probable error of the mean is $.6745 \frac{S.D.}{\sqrt{n}}$; of $S.D. = .6745 \frac{S.D.}{\sqrt{2n}}$.

FIGURE 68

THE AREA OF THE NORMAL CURVE, INSIDE (BLANK), AND OUTSIDE (SHADED), THE LIMITS SET BY ONE TIMES THE PROBABLE ERROR




## III. THE MEANING OF THE PROBABLE ERROR CONCEPT

The "significance" of individual measures and their means is measured in terms of their probable errors. The probable error is defined as a measure which added to and subtracted from the mean gives an amount within which the chances are even that an item selected at random will fall. It is said

[1] Pearl gives an interesting example of the variation of means taken from a series of random selections. Pearl, Raymond, *Introduction to Medical Biometry and Statistics*, Saunders, Philadelphia, 1923, pp. 210-213.

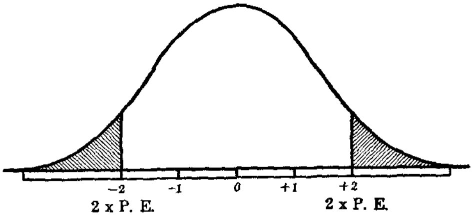
conventionally that if a certain result is three or more times as large as its probable error it is "significant." What is meant by this expression? The following illustrations taken from Pearl [1] will help to answer this question.

In Figure 68, the blank portion under the curve represents one half of the area. Accordingly, its boundaries on the $X$ axis mark the limits of the probable error.

In Figure 69, the corresponding blank portion representing twice the probable error comprehends 82.27 per cent of the area. The shaded portion includes 17.73 per cent of the area. Therefore, the odds are 82.27 to 17.73 or 4.64 to 1 that an item selected at random will fall within twice the probable error.

FIGURE 69

THE AREA OF THE NORMAL CURVE, INSIDE (BLANK), AND OUTSIDE (SHADED), THE LIMITS SET BY TWICE THE PROBABLE ERROR



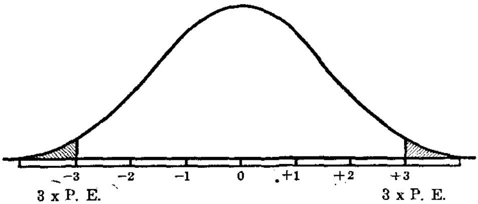

In Figure 70, the blank area is three times the probable error. It comprehends 95.70, while the shaded portions make up but 4.30 per cent of the total area. Therefore, the chances are 95.70 to 4.30 or 22.24 to 1, that an item taken at random will fall within three times its probable error. Similarly, the chances are 142.3 to 1 that an item will not exceed four times

[1] Pearl, Raymond, *Introduction to Medical Biometry and Statistics*, Saunders, Philadelphia, 1923, pp. 215-216.

its probable error. In this case the part of the total area of a probability surface falling outside of the limits of four times the probable error is less than 1 per cent—.698 per cent.[1]

FIGURE 70

THE AREA OF THE NORMAL CURVE, INSIDE (BLANK), AND OUTSIDE (SHADED), THE LIMITS SET BY THREE TIMES THE PROBABLE ERROR




To say that a measurement is "significant" when it is three or more times as large as its probable error is, therefore, equivalent to saying that the odds against its appearance—once in 22.24 times when three times the probable error is taken as a test—may be ignored. But as Pearl remarks: "As a matter of fact, this is not true, unless one chooses to regard 4.3 per cent as a negligible fraction of a quantity."[2]

The "odds" given above refer to the probable error of a single measure. Those for means, and standard deviations are different as indicated by the formulæ on p. 370. The probable error of a correlation coefficient is discussed later.[3]

## IV. SAMPLE MEASUREMENTS AND THE USES OF THE PROBABLE ERROR

Statistical studies are almost always made from samples. All prices cannot be included in computing an index number

[1] See Pearl, *op. cit.*, p. 218, for a table giving the "odds" for other relationships between a measurement and its probable error.

[2] *Op. cit.*, pp. 214-215.

[3] *Infra*, pp. 464-465.

nor all rents determined when studying family budgets. Neither the time required for all operators within manufacturing industries to complete an operation, nor the time necessary for every operator in telephone industries to answer the telephone calls of all subscribers, can be determined in order to answer specific inquiries. Samples must be used and some method employed for testing their reliability. Averages alone will not suffice; their limitations in describing frequency distributions have already been indicated. The most common measure of divergence from type is the standard deviation. But it is simply a measure for the samples taken. What is wanted is proof that the distribution in the samples indicates the distribution that would result if the whole "population" were included. The probable error supplies this. On the supposition that if all the population were included a distribution would follow the normal curve of error, the probable error stands in a mathematical relation to the standard deviation in the same way that the radius of a circle does to the circumference. Hence, the reliability of a sample may be expressed in terms of its probable error.

Breeders of animals and plants find it necessary to determine the probable error of their measurements in studies of variation from type.[1] Moreover, in the selection of men according to psychological and other tests,[2] in the grading of cotton and grains, in the setting of tasks, and the establishment of piece-rates of compensation on the basis of the "average" operator's performance, some measure of the reliability of the samples must be employed. Again, according to Fisher,[3] the only scientific method of establishing the pure premium for industrial accident insurance is to compare homo-

[1] Davenport, Eugene, *The Principles of Breeding*, Ginn and Co., New York, 1907, *passim*.

[2] Whipple, Guy M., *Manual of Mental and Physical Tests*, Warwick and York, Baltimore, Md., 1914, *passim*.

[3] Cf., Fisher, Arne, *Proceedings of the Casualty, Actuarial, and Statistical Society of America*, Vol. II, Part III, No. 6, May, 1916.

geneous conditions of risk exposure and to test the homogeneity by measures of dispersion in terms of their probable errors. Conformity to the normal law is proof that conditions are homogeneous. Most comparisons, it is held, involve non-homogeneous conditions. The proper unit is not the "establishment," but similar risk conditions in many establishments or industries.

It must be remembered that the probable error is a constant only for distributions of the normal probability form. It has no meaning for those which are markedly asymmetrical.

## V. Summary

The theory of probability and the properties of the normal law of error lie at the basis of most of the statistical studies of natural and physical phenomena. They have less application to problems growing out of human affairs where "chance" does not freely operate, and where measurements are not subjected to the law of error. Indeed, measurements of economic and business phenomena do not necessarily follow the probability form. They are generally asymmetrical or skewed. It is to the measurement of asymmetry or skewness to which we turn in the following chapter.

## References

Bowley, A. L., *Elements of Statistics* (4th Edition), King, London, 1920, pp. 259-286.

Fisher, Arne, *The Theory of Probability*, 2nd Edition greatly enlarged, Macmillan & Company, New York, 1922, *passim*.

Jones, D. C., *A First Course in Statistics*, Bell, London, 1921, Chapters XII, XIII and XVIII.

Kelley, T. L., *Statistical Method*, Macmillan & Company, New York, 1923, Chapter V.

Kent, F. C., *Elements of Statistics*, McGraw-Hill Book Co., New York, 1924, pp. 113-134.

Keynes, J. M., *A Treatise on Probability*, Macmillan & Company, London, 1921, *passim*.

Mills, Frederick C., *Statistical Methods Applied to Economics and Business*, Holt, New York, 1924, Chapter XV, pp. 516-547.
Rugg, H. O., *Statistical Methods Applied to Education*, Houghton Mifflin, Boston, 1917, Chapter VII.
Slichter, C. S., *Elementary Mathematical Analysis*, 2nd Edition, McGraw-Hill Book Company, New York, 1918, Chapter VII.
Whipple, G. C., *Vital Statistics*, Wiley & Sons, New York, 1919, Chapter XII.

# CHAPTER XII

## SKEWNESS OR ASYMMETRY

### I. Introduction

The preceding chapter was concerned with an elementary statement of the theory of probability and with the characteristics of the normal law of error curve or distribution which is expressive of this theory. While that which is probable must find its basis in experience, experience is finite and limited. Even the most protracted experiments of tossing coins, selecting cards from a pack, throwing dice, measuring the heights of soldiers, or the lengths of ears of corn have not succeeded in duplicating the probability curve which logic and belief prompt us to expect. All trials are limited in the sense that the entire "population" is not included and that time is not exhausted. Even though by repeated trials of coin tossing, for example, series secured by the expansion of the binomial were actually duplicated, such a result might be looked upon rather as an exception, the probability being almost certain that it would never be repeated.

The statistician deals with samples. His measurements are secured not under circumstances of pure chance, but under those peculiar to time, place, and particular environment. *Accordingly, the series which he selects do not exactly conform* to the probability curve.

An analogy at this place is in point. Perfect circles exist only in imagination. So also do the precise relations of their diameters and circumferences. Yet mathematicians are not debarred from drawing circles nor from using the constant, $\pi$ or 3.1416. So, likewise, pure probability distributions are

a creation of the imagination. Yet acknowledging this to be true, statisticians are not prevented from determining the degree to which distributions deviate from this ideal, nor from using the concept of probable error.

In the run of experience, statistical distributions are skewed or unsymmetrical. The purpose of this chapter is to describe the more important ways by which asymmetry or skewness may be measured.

## II. Dispersion and Skewness Contrasted

Measures and coefficients of dispersion, respectively, indicate absolutely and relatively the differences of the separate items in series from one taken as a standard. They measure deviations from type, varying emphasis being given to the differences depending upon the particular device used. The average deviation gives all of the differences their normal weight; the standard deviation accentuates those far removed from type. The quartile measure includes only those lying within the boundaries of the first and third quartiles. They do not, however, show the manner in which the deviations are distributed, nor do they localize them. They do not show the degree to which they cluster above or below the type selected.

Measures of skewness, on the other hand, indicate the position relative to the mode or median at which distributions are pulled away, distorted, or skewed from normality, i.e. from the symmetrical form of the curve of error. In the normal curve, mode, median, and arithmetic mean coincide. In unsymmetrical curves they differ in size. The function of measures of skewness is twofold: *(1) to indicate the direction of* skew or asymmetry, and (2) to measure the amount either absolutely or relatively.

Most, if not all distributions, are skewed to some degree,[1]

[1] Cf., Tolley, Howard R., "Frequency Curves of Climatic Phenomena," in *Monthly Weather Review*, United States Department of Agriculture, Vol. 44, November, 1916, pp. 634-642, 636.

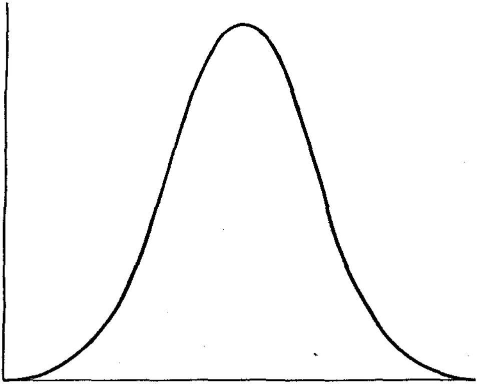
the *normal* distribution being in fact "abnormal" in the sense that it is never realized. Indeed, it is probably true that nature never repeats herself, although it is said that history does. Asymmetry in a particular case may be due among other things to imperfect measurements, inadequate sampling, personal bias, etc. In the universe at large, however, it probably rests upon more fundamental bases rooted in the fact of variation and diversity. But asymmetry takes a variety of forms—some marked, some slight—and it may be worth while briefly to illustrate certain of its types.[1]

## III. Types of Skewed Distributions

An ideal symmetrical frequency distribution is shown in Figure 71.

FIGURE 71

The Form of the Ideal Symmetrical Frequency Distribution



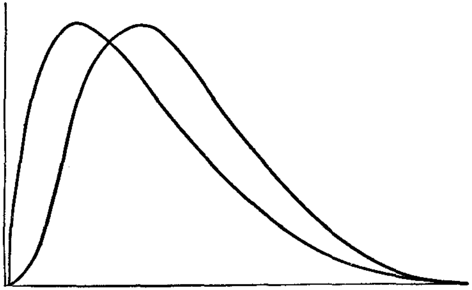
[1] More elaborate illustrations are given in Yule, G. U., *An Introduction to the Theory of Statistics*, Griffin, London, 1911, Chap. VI.

Two ideal distributions of the moderately asymmetrical type are shown in Figure 72.

A distribution approaching the moderately asymmetrical form is given in Table 29. Each of the curves in Figure 72 approaches the normal type—bell shaped—but neither is symmetrical. A mode is evident in each case but the items are not uniformly distributed about it—that is, distribution is skewed.

FIGURE 72

THE FORMS OF IDEAL MODERATELY ASYMMETRICAL OR SKEWED DISTRIBUTIONS



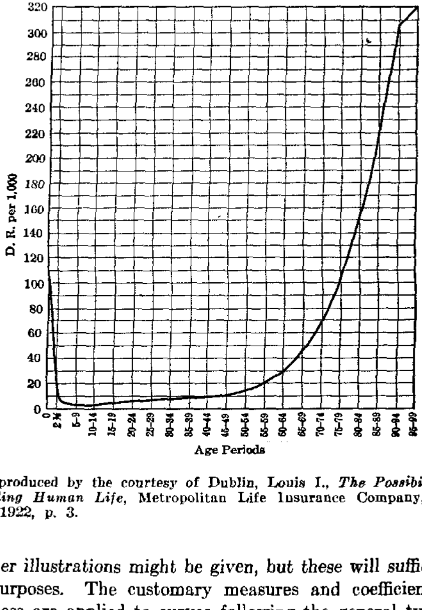
On the other hand, in Figure 73 the distribution is of quite a different kind[1]—the peculiar shape being primarily due to the fact that non-homogeneous groups in the attribute measured are grouped together.

Still another general type is encountered. Figure 74 shows an ideal J-shaped distribution, while four series approaching this form are given in the footnotes on pages 382 and 383.

[1] See Yule, *op. cit.*, p. 103, for an illustration of the ideal U-shaped form.

FIGURE 73

U-SHAPED DISTRIBUTION CURVE OF DEATHS PER 1,000 POPULATION AT SPECIFIED AGE PERIODS, UNITED STATES REGISTRATION STATES, 1920 *

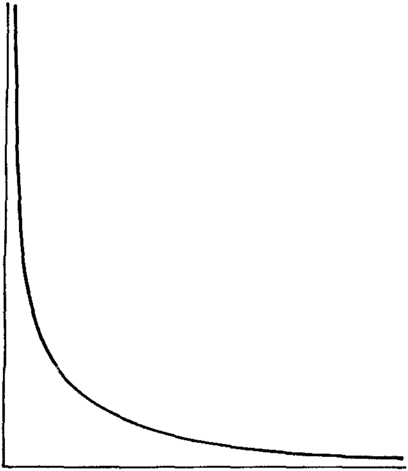



* Reproduced by the courtesy of Dublin, Louis I., *The Possibility of Extending Human Life*, Metropolitan Life Insurance Company, New York, 1922, p. 3.

Other illustrations might be given, but these will suffice for our purposes. The customary measures and coefficients of skewness are applied to curves following the general type of those in Figure 72—that is, where a mode is present, the distribution of items around it tending to be regular and systematic but where there is not a perfect balance on either

FIGURE 74

THE FORM OF THE IDEAL J-SHAPED FREQUENCY DISTRIBUTION CURVE



side. It is this type of curve with which we are concerned in the following section.

## IV. MEASURES AND COEFFICIENTS OF SKEWNESS

The chief and currently used measure of skewness is the difference between the arithmetic mean and the mode. If the mean exceeds the mode, skewness is said to be positive. If it is less than the mode, it is said to be negative. The mode,

*of course, is unaffected by extreme items whether large or* small. The arithmetic mean, on the other hand, is influenced not only by the size but also by the number of items. If distributions are normal, that is, if the "errors" in excess and in defect of the mean are equal in number and in extent of deviation, those which are positive cancel those which are negative, and the mean has the same position as the mode. If they are unsymmetrical, then the arithmetic mean may be greater or less than the mode depending upon the position of asymmetry. Accordingly, the difference between them may be used as a measure of skewness. The sign (+) or (—) secured by the computation, mean — mode, indicates the direction of skewness; the difference, indicates its amount.

Inasmuch as the mode as an average is not rigidly defined, its amount in a particular case may be in doubt. Interpola-

The following examples show distributions which are clearly asymmetrical:

*Illustration 1*

Number of Divorces in the U. S., 1887 to 1906, Classified by Number of Years of Married Life.
(*U. S. Statistical Abstract*, 1913, p. 85.)

| No. of Years Married | No. of Divorces |
|---|---|
| Total | 900,584 |
| Under 5 | 255,085 |
| 5 to 9 | 282,904 |
| 10 to 14 | 162,407 |
| 15 to 19 | 91,176 |
| 20 to 24 | 54,578 |
| 25 to 29 | 29,245 |
| 30 to 34 | 15,035 |
| 35 to 39 | 6,555 |
| 40 to 44 | 2,507 |
| 45 to 49 | 805 |
| 50 and over | 287 |

*Illustration 2*

Table Showing Number of Individuals and Corporations Assessed for Income Tax in 12 Wisconsin Counties, classified by amount groups of Assessed Incomes.
(*Rept. Wis. Tax Commission*, 1912, p. 37.)

| | |
|---|---|
| Total | 11,935 |
| Under $1000 | 7,890 |
| $1000 to $1999 | 1,910 |
| 2000 to 2999 | 786 |
| 3000 to 3999 | 406 |
| 4000 to 4999 | 234 |
| 5000 to 9999 | 411 * |
| 10,000 and over | 298 * |

* Notice the widths of the groups.

tion is then necessary. Various methods by which this may be done have already been suggested,[1] but each of them is more or less arbitrary. Different methods may give different amounts. But the above formula for skewness requires an exact mode—it cannot be used when the mode is given simply as a group or as falling within certain limits. A purely empirical interpolation formula for the mode, for moderately asymmetrical series, is as follows:

$$\text{Mode} = \text{mean} - 3\ (\text{mean} - \text{median})$$

That is, the median lies about one third of the distance from the mean toward the mode. This formula, however, should

Note continued from page 382.

*Illustration 3*

Table showing Distribution of Percentages of Cost of Collection to Total Collections, Internal Revenue of the U. S., 67 Districts, 1913. (Compiled from the *Report of the Commissioner of Internal Revenue*, 1913, p. 211.)

| PERCENTAGE GROUPS | NO. OF DISTRICTS (Frequency) |
|---|---|
| TOTAL | 67 |
| 0 to 2 | 29 |
| 2 to 4 | 24 |
| 4 to 6 | 4 |
| 6 to 8 | 4 |
| 8 to 10 | 4 |
| 10 to 12 | 0 |
| 12 to 14 | 1 |
| 14 to 16 | 1 |

*Illustration 4*

Number of Weavers weaving Worsted Goods in the U. S. and Receiving Specified Wage-rates Based upon Actual Weaving Time on Yardage at Regular Piece-rates per Yard, Including Ordinary Stoppage of Loom. (*Report of Tariff Board on Schedule K*—Vol. IV, p. 1007.)

| EARNINGS PER HOUR | NUMBER |
|---|---|
| TOTAL | 3182 |
| 10 to 12 | 165 |
| 12 to 14 | 275 |
| 14 to 16 | 375 |
| 16 to 18 | 490 |
| 18 to 20 | 490 |
| 20 to 22 | 438 |
| 22 to 24 | 414 |
| 24 to 26 | 235 |
| 26 to 28 | 150 |
| 28 to 30 | 108 |
| 30 to 32 | 34 |
| 32 to 34 | 4 |
| 34 or over | 4 |

[1] See Chapter IX, pp. 297-307.

be used with caution. It does not hold for markedly asymmetrical distributions because of the effect which exceptionally small or large items have on the mean. The fact of skewness may be determined by rough methods—even by inspection in most cases—but a measurement of the *degree* of skewness by this method necessitates the location of an exact mode.

But more than a measure of skewness is required if series in this respect are to be compared. Differences between means and modes, as the amounts themselves, are always expressed in the unit in which series are measured. These may be feet, inches, gallons, dollars, cents, or what not. It is meaningless, therefore, to say that because the difference between the mean and mode in one series expressed in dollars, for instance, is larger than the difference in another series expressed in cents, feet, or inches, that skewness or asymmetry is greater. Some method of reducing the amounts to a common denominator must be used before comparison is possible. It is asymmetry which is being compared; not the units in which the measurements are made. What common denominator is most suitable?

Skewness is divergence from symmetry, and symmetry is uniform dispersion with respect to the mean. Standard and average deviations for series which are widely dispersed are large; for those which are narrowly dispersed, they are small. The most satisfactory measure of dispersion being the standard deviation, or *S.D.*, this may be used as a divisor in order to reduce to the same denomination amounts of skewness. Accordingly, the *coefficient* of skewness based on the positions of the mean and the mode is

$$\frac{Mean - Mode}{S.D.}$$

The measurement of skewness is always indicated by a plus (+) or minus (—) sign prefixed to an amount, the unit being the same as that in which a series is measured. The coefficient of skewness is always indicated by a decimal pre-

fixed by a plus (+) or minus (—) sign, the units in the numerator and in the denominator in the formula, $\frac{mean-mode}{S.D.}$, being the same. When the mean and the mode coincide, both the measure and the coefficient are zero.

But the position, measure, and coefficient of skewness may be secured for a part rather than for the whole of a series. The conventional method is to measure the portion lying between the first and the third quartiles. If a series is symmetrical for this half, the quartiles are equally distant from the median. That is, one half of the difference between them when added to the lower or subtracted from the upper quartile gives the median amount. Accordingly, the nature and amount of skewness, within the quartile range, is indicated by the formula

$$(Q^3 + Q^1) - 2\ (Median)$$

If the quartiles are equally distant from the median, this formula gives zero. If the distance from the median to the upper quartile exceeds that from the lower quartile to the median, the formula gives a positive quantity. If the reverse is true, it gives a negative amount. The *position* of skewness —that is, relative to the median but applying only to the middle half of a series—is indicated by the nature of the sign. The *amount* of skewness is shown by the quantity accompanying the sign.

But this measure like that based upon the mean and the mode must be stated as a ratio before comparisons can be made between series measured in different units. An appropriate common denominator is $Q^3 - Q^1$. The formula for the coefficient of skewness based on the quartiles is, therefore, as follows:

$$\frac{(Q^3 + Q^1) - 2(Median)}{Q^3 - Q^1}$$

the result being written with a plus (+) or minus (—) sign as a prefix.

One half the total frequencies are included between $Q^1$ and $Q^3$. In a symmetrical distribution, $Q^3$ and $Q^1$ are equally distant from the median. In an asymmetrical distribution, this *is not the case. If for this part of the series skewness is positive,* the third quartile is farther removed from the median than is the first quartile. If skewness is negative, the reverse is true.

The quartile *type* of skewness measure may also be applied to the halves of series above and below the median. If it is applied to the lower half, the formula is

$$\frac{\textit{Smallest item} + \textit{Median} - 2\ (Q^1)}{\textit{Median} - \textit{Smallest item}}$$

If it is applied to the upper half, the corresponding formula is

$$\frac{\textit{Largest item} + \textit{Median} - 2(Q^3)}{\textit{Largest item} - \textit{Median}}$$

A positive or negative measurement or coefficient of skewness of a series shows that it is not normal. In the measure and coefficient based on the mean and the mode, asymmetry is localized relative to the mode. In those based on the quartiles, it is indicated relative to the median. But the median and the mode are identical in normal distributions. In those which are skewed, the mode is least and the arithmetic mean most affected by asymmetry. The median holds an intermediate position.

## V. Methods of Summarizing Frequency Series

The three primary methods of summarizing frequency series are (1) to average the gross items using the arithmetic mean, median, mode, or other suitable measure; (2) to summarize by the method of averages or otherwise the deviations (errors) of the items from a standard or type—that is, to calculate

measures and coefficients of dispersion; (3) to determine the *nature and amount*, if *any*, of skewness, that is, departure from the symmetry of the normal probability distribution.

An adequate description of a statistical series requires not alone one of these summaries but all of them. Each of them tells a different story. If the averages of gross items closely agree, the normal law of error distribution is approached; if dispersion is small, the measures tend to be homogeneous. If skewness is present and negative large deviations are found below the mean; if it is present and positive, such deviations are above the mean.

Measures and coefficients of both dispersion and skewness should be in everyday use in statistical work. For two or more series arithmetic means may be identical, but dispersion and skewness different. These facts are important. Current comparisons of sales, wages, interest rates, stock and bond prices, etc., by means of such measures could not fail to throw new light on the problems of business.

Without carrying through the arithmetical steps in the computation of different summaries—since this would involve unnecessary repetition of the methods already given—their use may be illustrated by comparing wage data for a single occupation in eighteen identical establishments, reported by the United States Bureau of Labor Statistics.[1]

Table 64 gives the classified wage data and the summaries computed from them. Figures 75 and 76 show graphically the detail of the series and the positions at which the different averages fall.

What are some of the things which these summary figures show?

1. The arithmetic mean exceeds both the median and the mode[2] in each year. Skewness is, therefore, positive.

[1] *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 190, May, 1916, p. 139.

[2] A single mode is indeterminate in 1908 and 1910.

TABLE 64

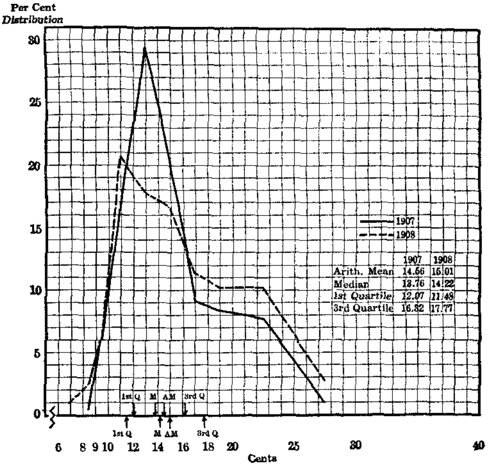
TABLE SHOWING CLASSIFIED WAGE-RATES OF FEMALE MENDERS IN EIGHTEEN IDENTICAL WOOLEN AND WORSTED MANUFACTURING ESTABLISHMENTS, BY YEARS, TOGETHER WITH CERTAIN MEASURES OF DISPERSION * AND SKEWNESS *

| WAGE GROUPS—CENTS PER HOUR | CLASSIFIED WAGE-RATES OF FEMALE MENDERS, BY YEARS | | | |
|---|---|---|---|---|
| | 1907 | 1908 | 1909 | 1910 |
| Total | 403 | 341 | 583 | 498 |
| 6 to 8 | — | 3 | 3 | 1 |
| † 8 to 9 | 2 | 8 | 44 | 14 |
| † 9 to 10 | 27 | 22 | 91 | 44 |
| 10 to 12 | 68 | 71 | 117 | 125 |
| 12 to 14 | 119 | 61 | 82 | 81 |
| 14 to 16 | 81 | 57 | 86 | 58 |
| 16 to 18 | 37 | 39 | 49 | 30 |
| 18 to 20 | 34 | 35 | 42 | 82 |
| † 20 to 25 | 31 | 35 | 58 | 43 |
| 25 to 30 | 4 | 10 | 11 | 16 |
| † 30 to 40 | | | | 4 |
| ¶ 40 and over | | | | |
| Arithmetic Mean | 14.56¢ | 15.01¢ | 13.96¢ | 14.97¢ |
| Mode (by interpolation) | 13.08¢ | ** | 10.95¢ | ** |
| First Quartile | 12.07¢ | 11.48¢ | 10.14¢ | 11.05¢ |
| Median (Second Quartile) | 13.76¢ | 14.22¢ | 12.09¢ | 13.62¢ |
| Third Quartile | 16.32¢ | 17.77¢ | 16.61¢ | 18.52¢ |
| Dispersion: | | | | |
| Average Deviation | 2.86¢ | 3.54¢ | 3.75¢ | 4.00¢ |
| Standard Deviation | 3.67¢ | 4.47¢ | 4.58¢ | 4.96¢ |
| Coefficient on A. D. | .196 | .236 | .269 | .287 |
| Coefficient on S. D. | .252 | .298 | .328 | .331 |
| Skewness: | | | | |
| Arithmetic Mean—Mode | +1.48¢ | ** | +3.01¢ | ** |
| Quartile Measure | +.87¢ | +.81¢ | +2.57¢ | +2.33¢ |
| Coefficient on S. D. | +.40 | ** | +.66 | ** |
| Coefficient on Quartile | +.21 | +.13 | +.40 | +.31 |

* Computed.
† Notice size of group.
¶ Notice residuum.
** Indeterminate.

FIGURE 75

CURVES SHOWING, BY YEARS, CLASSIFIED WAGE-RATES OF FEMALE MENDERS IN WOOLEN AND WORSTED ESTABLISHMENTS, 1907-1908



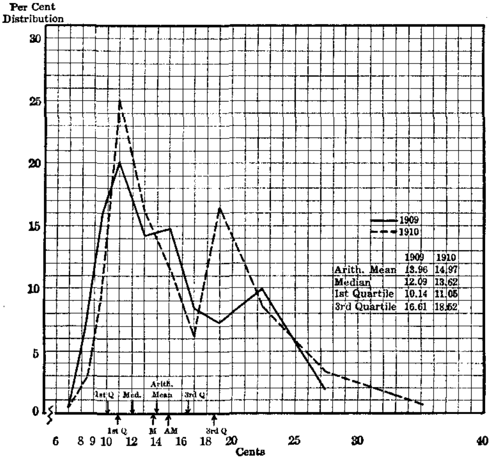

2. Both the average and the standard deviations, as well as the coefficients of dispersion based on them, tend to increase from year to year. That is, the average differences in rates when measured from the arithmetic mean tend to be larger both absolutely and relatively.

3. The lower quartile position in 1907 is essentially as high as the median in 1909. The range of difference in rates between the median and the upper quartile is more than double in 1910 what it is in 1907.

4. In both 1909 and 1910 there is a much more pronounced

FIGURE 76

CURVES SHOWING, BY YEARS, CLASSIFIED WAGE-RATES OF FEMALE MENDERS IN WOOLEN AND WORSTED ESTABLISHMENTS, 1909-1910




skew between the medians and the upper quartiles than in 1907 and 1908, the coefficients on the quartile measures being, respectively, + .21, + .13, + .40, + .31.

5. The wage-rates which the middle-half received varied as follows:

1907, from 12.07 to 16.32 or 4.25¢.
*1908*, from *11.48* to *17.77* or 6.29¢.
1909, from 10.14 to 16.61 or 6.47¢.
1910, from 11.05 to 18.52 or 7.47¢.

That is, the position of the lower quartile, with one exception, has fallen, and that of the upper quartile, with one ex-

ception, risen. While the average rate in 1910 is less than one half cent higher than in 1907, the wage of the person three fourths up in the scale is more than two cents higher.

6. The coefficient of dispersion based on the average deviation, and the coefficient of skewness based on the quartile measure are higher in 1909 and 1910 than in any other of the years. Negative skewness indicates a healthy influence in wage conditions—a concentration above the arithmetic mean. On the other hand, the wide absolute and relative dispersions tend to counteract this.

Other detailed facts may be gleaned from a comparison of these summaries, but those given are sufficient to show how they may be used.

## VI. Conclusion

It is generally not enough to speak in terms of averages when characterizing statistical series. Deviations both as to amount and position are frequently quite as important as the averages themselves. After all, any sort of summary sacrifices part of the detail; but the sacrifice is less when different types are used to supplement each other than when reliance is placed in one alone.[1]

## References [2]

Bowley, A. L., *Elements of Statistics*, 4th Edition, King, London, 1920, pp. 116-117.

Jones, D. Caradog, *A First Course in Statistics*, Bell, London, 1921, pp. 61-68.

[1] See Secrist, Horace, "Competition in the Retail Distribution of Clothing—a Study of Expense or 'Supply' Curves," *Bureau of Business Research, Northwestern University*, Series II, Number 8, Chicago, 1923, where, as a test of the adequacy of sample data, the ideal relations between the average and the standard deviations, and the standard deviation and the probable error are applied to the data used. In this particular case the samples closely conformed to the normal curve of error, thus giving evidence that the sample was representative of the total "population."

[2] See references to Chapter X, p. 359.

KING, W. I., *Elements of Statistical Method*, Macmillan, New York, 1912, pp. 159-166.

MITCHELL, W. C., "The Characteristics of Price Fluctuations," in *Index Numbers of Wholesale Prices in the United States and Foreign Countries*, Washington, D. C., 1921, pp. 11-23.

YULE, G. UDNY, *An Introduction to the Theory of Statistics*, 3rd Edition, Griffin, London, 1916, pp. 149-153.

# CHAPTER XIII

## THE THEORY AND MEASUREMENT OF CORRELATION

### I. Introduction

Any body of data or any statistical series may be analyzed descriptively by giving the details in tabular or in graphic form. If summaries are appropriate, averages of different types may be taken of the gross items and the relations of one to the other indicated. With these statistical abbreviations as points of departure, the deviations of the gross items may be further summarized by the use of averages. That is, dispersion in its absolute and relative aspects may be computed. But since dispersion indicates neither symmetry nor divergence from normal, measures and coefficients of skewness are required.

If two or more bodies of data or statistical series are to be compared, any one or all of these devices may be used. Tabular and graphic forms give the detail; averages, when expressed in a common unit, admit of direct comparison. For instance, a statement such as the following is significant: The average expense of doing business in retail meat stores is 19 per cent; in retail clothing stores, 24 per cent of sales. On the other hand, statements such as these have no comparative meaning: the average rent expense is 2 per cent of sales in retail meat stores; in the same type of stores the average number of times stock is turned is once in two days. Both amounts are averages, but not of the same things. Hence, comparatively, they are meaningless.

Moreover, the amounts of dispersion in two series cannot

be compared by means of their standard deviations unless the averages from which they are computed are identical. If they are different, ratios are required, the respective averages constituting a common denominator with which to reduce the absolute amounts to a relative basis. The same type of observation applies to measures of skewness. It is impossible to compare degrees of asymmetry in two or more series by saying that in one skewness is + 7 and in another + 4. The 7 and 4 have comparative significance only in case the standard deviations are identical. If they are not the same, the respective standard deviations as divisors reduce them to the same denomination. Comparison is then possible.

Now in all statistical work comparison of one sort or another is the goal. In some cases what is wanted are comparative pictures of a single series as shown by *different* measures of its attributes.[1] In others, it is a comparative picture of different series by the use of the *same* measures of their properties.[2]

But not infrequently one desires to compare for two or more series the corresponding deviations from their respective averages. That is, interest lies in getting a statistical measure of congruence of change in the deviations. In this case, pairs of values are dealt with, the purpose being to measure the manner and degree in which they *concurrently* fluctuate or deviate from a norm or standard. A ratio of some sort which will summarize the relations which they bear to each other is needed.

## II. Comparison, Causation, and Correlation

Comparison can be made only between things possessing common qualities. These may be of time, of place, or of condition. For instance, the accident rate in a given industry may be compared before and after the installation of safety

[1] *See the different summaries in any one of the columns in Table 64.*
[2] See the corresponding summaries in the different columns in Table 64.

devices. Moreover, comparisons may extend to two industries operating at different places or under different conditions, the purpose being merely to record a quantitative difference. But they are rarely made for this end alone. Generally, a more or less definite purpose of establishing a causal connection lies in the background. A specific inquiry is undertaken to determine whether phenomena stand in the relation of cause and effect, or whether they are the result of a common cause.

To establish cause and effect relations between economic and social phenomena, however, is as alluring as it is difficult. Such phenomena grow out of the facts of business, the observations of science, the records of history, etc., and are interpreted differently by different people, at different times, and for different conditions. Their seeming unity and identity are only relative, and the order of cause and effect not hard and fast.

Variations at a given time and changes over a period of time, characteristic of our economic and social life, are all traceable to a complex of causes.[1] A given cause is not homogeneous except when viewed in the most superficial manner. Moreover, its "effects" are not always the same; they vary. In some cases "cause and effect" seem to be coincident in time; in others the "effects" follow the "causes" as sequences spread over long or short periods. Indeed, what appears to be a "cause" may be an "effect" of an antecedent "cause." In the physical, natural, and social world, "cause and effect" are in reality variates.[2] How true this is may be seen by briefly referring to some of the more common relations among business phenomena.

Stimulation of business shows itself in an increase of bank debits, but not all banks are equally affected. Interest rates ultimately respond but not uniformly in different markets. Excessive issues of irredeemable paper currency ultimately

[1] See the definition of Statistics, *supra*, p. 10 *ff*.
[2] *Cf.* Hooker, R. H., "Correlation of the Marriage Rate with Trade," *Journal of the Royal Statistical Society*, Vol. 64, p. 485.

result in a premium on gold and in a general increase in prices, but not concurrently with the issue nor to the same degree for different types of business transactions. The surplus reserves of banks are said currently to fix the call-loan interest rate. But not all loans, nor all banks nor customers are affected at the same time and to the same degree. Wholesale and retail prices fluctuate together, but the former fall first and rise first, the latter following some distance behind. The effect of cotton prices on acreage is shown only from one cropping to another, and then not uniformly over the cotton area. Wages undoubtedly tend to rise with rising prices, but *not coincidently, nor to the same degree in all trades.* Business prosperity undoubtedly stimulates immigration but only after a period of time. The relation is sequential. Moreover, *general* prosperity is far from uniform for areas, for industries, and for classes.[1]

Comparison, therefore, involves pairing things or events which are not identical in all particulars as to time, place, and condition. Causation in fact becomes correlation. A study of cause and effect, whether of coincidence or sequence, becomes largely a study of association. The idea that a given effect is the result of a specific cause, or that the effect must in the nature of the case be uniform and absolute, does not apply to business and economic phenomena. Causes never operate under exactly the same circumstances. Oneness of effect is only apparent, variation being evident the moment that the scale of measurement is reduced.[2]

Business does not go on indefinitely repeating itself in one unending round of sameness. Variation characterizes all phenomena which involve the human element, whether viewed as

[1] See King, W. I., ***Employment Hours and Earnings in Prosperity and Depression, United States, 1920-1922***, The National Bureau of Economic Research, New York, 1923, ***passim.***

[2] When making comparison in economics or business, there is a tendency to attempt to safeguard oneself against error and criticism by introducing the proviso—*other things being equal.* But the "other things" are rarely if ever equal in actual life.

cause or as effect. The tendency to look upon business and economic phenomena in a mechanistic manner, to expect a complete and narrow fulfilment of *the law* of cause and effect, needs to be dispelled. Just as soon as it is, the way is open for the use of scientific method. This is the method of discrimination, of the study of small differences, of acting in the light of facts properly interpreted, and of reducing them as classified knowledge into rules of action.

The conclusion to which facts point may be nothing more, for instance, than that it is unwise to market corn with high moisture content, since weight varies inversely with moisture,[1] or to leave corn in leaky cars exposed to hot weather because both are conducive to the development of acidity, and acidity retards germination;[2] that a "bacon" hog can be produced; that corn grown from seed from ears 10 inches long has, on the average, longer ears than corn grown from seed of ears that are 8 inches long;[3] that the prices of bonds with fixed interest rates vary inversely with general commodity price changes;[4] that a farm of less than 40 acres in a certain district is economically undesirable;[5] that the milk production of cows increases until the animals are at least six years of age and then falls off;[6] that there is a direct relation

[1] *Bulletin* of the United States Department of Agriculture, No. 472, October, 1916, "Improved Apparatus for Determining the Test Weight of Grain, with a Standard Method of Making the Test." See curve on p. 4.

[2] *Bulletin* of the United States Department of Agriculture, No. 102, July, 1914, on "Acidity as a Factor in Determining the Degree of Soundness of Corn," pp. 12, 14, *passim*.

[3] "Type and Variability in Corn," *Bulletin* No. 119, University of Illinois Agricultural Experiment Station, October, 1907.

[4] Mitchell, Wesley C., *Business Cycles*, University of California Studies, Berkeley, 1913, pp. 201-219, especially charts 23 and 24, pp. 206 and 207, respectively.

[5] *Bulletin* of the United States Department of Agriculture, No. 341, January, 1916, on "Farm Management Practice of Chester County, Pa.," pp. 56 ff.

[6] Holdaway, C. W., "Statistical Weighting for Age of Advanced Registry Cows," *The American Naturalist*, Vol. 50, No. 559, p. 681.

between fatigue and industrial accidents;[1] that accident rates tend to increase with expanding and to contract with falling business;[2] that twin offspring from twin parents in sheep production is more common than from parentage conforming to any other condition;[3] etc. Whatever they are and to whatever type of business they apply, if they are arrived at as a result of a dispassionate study of facts in an attempt to determine association and correlation and not to *prove* the infallibility of some narrow cause-and-effect relationship, a clear advance is made in the use of statistical methods.

## III. The Meaning of Correlation

### 1. DEFINITION AND EXPLANATION

If it is impossible in social affairs to establish causation in a narrow sense, since causes operate as variations and effects show themselves in the same way, it is unnecessary to conclude that cause-and-effect relations in a larger sense cannot be measured. The problems are different. The first is the impossible task of establishing an absolute cause and an absolute effect; the latter is the problem of measuring correlation. Pearson makes the distinction clear in the following passage:

"When we vary the cause, the phenomenon changes, but not always to the same extent; it changes, but has variation in its change. The less the variation in that change, the more nearly the cause defines

[1] "The Case of the Shorter Day," *Franklin O. Bunting vs. The State of Oregon*, Brief for the Defendant in Error, by Felix Frankfurter, Vol. I, pp. 165-193.

[2] Mowbray, A. H., and Black, S. B., "Relation of Accident Frequency to Business Activity," in *Proceedings of the Casualty, Actuarial and Statistical Society of America*, Vol. II, Pt. III, No. 6, May, 1916, pp. 418-426.

[3] Rietz, H. L., and Roberts, Elmer, "Degree of Resemblance of Parents and Offspring with Respect to Birth of Twins for Registered Shropshire Sheep," in *Journal of Agricultural Research*, Vol. IV, No. 6, September, 1915.

the phenomena, the more closely we assert the association or the correlation to be. It is this conception of correlation between two occurrences, embracing all relationships from absolute independence to complete dependence, which is the wider category by which we have to replace the old idea of causation. Everything in the universe occurs but once, there is no complete sameness of repetition. Individual phenomena can only be classified, and our problem turns on how far a group or class of like, but not absolutely same, things which we term 'causes' will be accompanied or followed by another group or class of like, but not absolutely same, things which we term 'effects.' "[1]

What correlation, as thus distinguished from causation, means is indicated by Davenport as follows:

"The whole subject of correlation refers to that interrelation between *separate characters by which they tend, in some degree,* at least, to move together. This relation is expressed in the form of a ratio. Thus, if an increase of one character is always followed by a corresponding and proportional increase in a related character, the correlation is said to be perfect and the ratio is 1. On the other hand, if an increase in one character is followed by a corresponding and proportional *decrease* in a related character, the correlation is said to be negative and the ratio is —1, or perfect negative correlation. Still again, if the characters in question are absolutely indifferent the one to the other, the correlation is said to be zero, indicating mere association under the law of independent probability, without causative relation of any kind."[2]

Probability, as briefly described in Chapter XI, was said to supply a basis for the theory of error. Under conditions of pure chance, frequency measurements describe the normal law of error curve. The basis for expecting such curves is found in games of chance such as coin tossing, selection of balls from an urn, etc. In spite of the fact that such distributions are ideal and probably never realized in actual experience, they are the basis for much of our statistical reasoning.

Back of the theory of error and of normal distributions rests

[1] Pearson, Karl, *The Grammar of Science*, 3d Edition, Black, London, 1911, p. 157.

[2] Davenport, Eugene, *Principles of Breeding*, Ginn & Company, New York, 1907, p. 453.

the assumption that chance freely operates—that is, that every condition is the result of a multitude of causes, all operating to produce an effect, but independent of each other. Accordingly, "causes" and "effects" are characterized by variation.

## 2. ILLUSTRATIONS OF CORRELATION BY THROWS OF DICE

Darbishire,[1] by throwing 12 dice 1000 times and counting the number at each throw which had four or more spots uppermost,[2] secured the results shown in Table 65.

TABLE 65

TABLE SHOWING THE DISTRIBUTION OF DICE WITH FOUR OR MORE SPOTS UPPERMOST IN 1000 THROWS

| Result of Throw | Frequency | Result of Throw | Frequency |
|---|---|---|---|
| 0 | 0 | 7 | 179 |
| 1 | 3 | 8 | 129 |
| 2 | 15 | 9 | 64 |
| 3 | 55 | 10 | 11 |
| 4 | 110 | 11 | 2 |
| 5 | 208 | 12 | 1 |
| 6 | 223 | | |

That is, chance operating freely produced a distribution closely approaching the normal type. The significant thing, however, is that it is not perfectly normal. If another set of 1000 trials of the same kind were made, a similar approximation to normal distribution would be secured. The probability is almost certain, however, that the results in the second case

[1] Darbishire, A. D., "Some Tables for Illustrating Statistical Correlation," in *Memoirs and Proceedings of the Manchester Literary and Philosophical Society*, Vol. 51, No. 16, 1907. This is in continuation of a similar study made by Weldon, W. F. R.—"Inheritance in Animals and Plants," pp. 81-100, in *Lectures on the Method of Science*, edited by T. B. Strong, Oxford, 1906.

[2] The probability that *any* side of a perfect cube if thrown will come up is equal to that of any other side. The probability that a certain side will come up is 1/6.

would not be exactly the same as those in the first.[1] That is, the "causes" give varying "effects."

Successive throws, after each of which all dice are returned to the receptacle and thrown again, are entirely distinct. There is no connecting link between them which makes them stand in the relation of cause and effect. The different sets of trials and each throw in each trial are independent.

If two such trials of 500 throws each are tabulated so that the result in each first throw is paired with that in each second throw, the detail of Table 66 is secured. This is a double frequency table, provision being made in the stub for recording the results in the first throws, and in the caption, for the results in the second throws.

TABLE 66

TABLE GIVING THE RESULTS OF 500 PAIRS OF THROWS OF 12 DICE WHEN ALL THOSE THROWN THE FIRST TIME WERE THROWN THE SECOND TIME *

| | | | Second Throws | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | | Total | → | 1 | 9 | 24 | 57 | 112 | 101 | 94 | 62 | 31 | 6 | 2 | 1 |
| | 0 | ↓ | | | | | | | | | | | | | |
| | 1 | 2 | — | — | — | — | — | 1 | — | — | 1 | — | — | — | — |
| | 2 | 6 | — | — | — | 1 | — | 4 | — | — | — | 1 | — | — | — |
| | 3 | 31 | — | — | — | 1 | 4 | 7 | 8 | 5 | 4 | 1 | 1 | — | — |
| | 4 | 52 | — | — | 4 | 4 | 7 | 9 | 6 | 12 | 5 | 5 | — | — | — |
| First Throws | 5 | 95 | — | — | 3 | 5 | 13 | 26 | 14 | 14 | 12 | 6 | 1 | 1 | — |
| | 6 | 123 | — | — | 1 | 6 | 15 | 25 | 24 | 28 | 15 | 6 | 2 | 1 | — |
| | 7 | 87 | — | — | 1 | 5 | 7 | 16 | 22 | 15 | 13 | 6 | 1 | — | 1 |
| | 8 | 66 | — | — | — | 1 | 7 | 15 | 19 | 12 | 6 | 6 | — | — | — |
| | 9 | 33 | — | 1 | — | 1 | 2 | 9 | 7 | 6 | 6 | — | 1 | — | — |
| | 10 | 5 | — | — | — | — | 2 | — | 1 | 2 | — | — | — | — | — |
| | 11 | | — | — | — | — | — | — | — | — | — | — | — | — | — |
| | 12 | | — | — | — | — | — | — | — | — | — | — | — | — | — |

* The order of the units on the ordinate scale is reversed in this instance from that usually followed.

An inspection of the table shows little or no connection between the results secured in the first and in the second

[1] *Cf.* Weldon, W. F. R., *op. cit.*, for the results of three trials.

throws of each trial. For each of the precise results in the first throws there is a variety of results in the second. Similarly, for each of the precise results in the second throws there is a variety of results in the first. For instance, when there are 7 dice in the first throws with 4 or more spots uppermost, there are from 2 to 12 with 4 or more in the second trials. Dispersion is equally noticeable in the opposite direction. When 8 are secured in the second trials, the corresponding numbers in the first throws vary from 1 to 9.

The totals for the first as well as for the second throws give close approximations to the normal curve. The most probable number of dice showing 4 or more spots uppermost in a throw of twelve is six, but the number may be anything between zero and 12. The concentration at or near six in the totals and in the arrays—distributions in lines and columns—shows this to be true.

TABLE 67

TABLE GIVING THE RESULTS OF 500 CONNECTED THROWS OF 12 DICE, IN EACH SECOND THROW OF WHICH 3 DICE WERE LEFT DOWN AND COUNTED *

| | | | Second Throws | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | | Total | → | 2 | 7 | 31 | 55 | 82 | 111 | 108 | 71 | 25 | 7 | 1 | — |
| | 0 | ↓ | | | | | | | | | | | | | |
| | 1 | 2 | — | — | — | — | 1 | — | 1 | — | — | — | — | — | — |
| | 2 | 7 | — | — | — | — | — | — | 6 | 1 | — | — | — | — | — |
| | 3 | 20 | — | 1 | 1 | 5 | 2 | 2 | 4 | 5 | — | — | — | — | — |
| | 4 | 64 | — | — | 1 | 8 | 6 | 21 | 16 | 6 | 6 | — | — | — | — |
| First Throws | 5 | 92 | — | — | 4 | 3 | 12 | 15 | 23 | 22 | 9 | 3 | 1 | — | — |
| | 6 | 123 | — | 1 | — | 10 | 16 | 17 | 23 | 28 | 22 | 5 | 1 | — | — |
| | 7 | 97 | — | — | 1 | 4 | 9 | 17 | 18 | 24 | 16 | 5 | 3 | — | — |
| | 8 | 54 | — | — | — | 1 | 5 | 6 | 10 | 14 | 8 | 7 | 2 | 1 | — |
| | 9 | 30 | — | — | — | — | 4 | 3 | 9 | 6 | 6 | 2 | — | — | — |
| | 10 | 10 | — | — | — | — | — | 1 | 1 | 1 | 4 | 3 | — | — | — |
| | 11 | 1 | — | — | — | — | — | — | — | 1 | — | — | — | — | — |
| | 12 | | — | — | — | — | — | — | — | — | — | — | — | — | — |

* See note to Table 66.

Independent forces in a universe of chance gave the results in Table 66. But the "chance" distribution in the first throws may be made to determine (cause) those in second throws.

Such "causation" was accomplished by Darbishire as follows: In order to connect or relate the two throws of each pair, he repeated the experiment, first leaving down and counting in the second throw of each pair one, then two, then three, etc., of the dice which previously had been stained red so as to distinguish them from the others. The experiment was continued until all of the 12 dice thrown in the first, were left down for the second throws. The results when 3, 5, and 10 dice were left down are given in Tables 67, 68, and 69, respectively. A graphic picture of the dispersion of the throws is shown in Figure 77.

In each pair of trial throws, in which one or more of the dice is left on the board and counted in the second throw, there is a common element. That is, the first is in part a cause of

TABLE 68

TABLE GIVING THE RESULTS OF 500 CONNECTED THROWS OF 12 DICE, IN EACH SECOND THROW OF WHICH 5 DICE WERE LEFT DOWN AND COUNTED *

| | | | SECOND THROWS | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | | Total | → | | 11 | 20 | 54 | 93 | 112 | 118 | 60 | 21 | 9 | 2 | — |
| | 0 | ↓ | | | | | | | | | | | | | |
| | 1 | 2 | — | — | — | — | 1 | 1 | — | — | — | — | — | — | — |
| | 2 | 11 | — | — | 3 | 1 | 5 | 1 | 1 | — | — | — | — | — | — |
| | 3 | 26 | — | — | 3 | 3 | 8 | 4 | 4 | 4 | — | — | — | — | — |
| | 4 | 69 | — | — | 3 | 6 | 9 | 21 | 14 | 10 | 5 | 1 | — | — | — |
| First | 5 | 83 | — | — | — | 4 | 11 | 23 | 21 | 15 | 9 | — | — | — | — |
| Throws | 6 | 109 | — | — | 1 | 3 | 9 | 18 | 27 | 29 | 16 | 3 | 2 | 1 | — |
| | 7 | 95 | — | — | 1 | 2 | 5 | 14 | 24 | 28 | 10 | 7 | 4 | — | — |
| | 8 | 63 | — | — | — | 1 | 5 | 9 | 10 | 18 | 14 | 4 | 2 | — | — |
| | 9 | 31 | — | — | — | — | — | 2 | 9 | 13 | 4 | 3 | — | — | — |
| | 10 | 10 | — | — | — | — | 1 | — | 2 | — | 2 | 3 | 1 | 1 | — |
| | 11 | 1 | — | — | — | — | — | — | — | 1 | — | — | — | — | — |
| | 12 | | — | — | — | — | — | — | — | — | — | — | — | — | — |

* See note to Table 66.

the second, exerting an influence in proportion to its size. But the distributions in none of the cases, if the trials were repeated, would necessarily follow the order here given. Causes never operate at different times under exactly the same conditions, and the effects that follow from them are not always and necessarily the same. To duplicate the conditions under which causes operate will not necessarily duplicate the effects. "Duplication" after all in any way except as approximation is impossible in actual life.

How nearly economic and business phenomena remain homogeneous for any appreciable period, even in an approximate sense, is always doubtful. The forces affecting them are always in a state of flux governed as they are by population composition, state of trade, distribution of wealth, custom, fad, fashion, prejudice, etc. The whole range of human reaction is exhibited in more or less degree. Statistics under

TABLE 69

TABLE GIVING THE RESULTS OF 500 CONNECTED THROWS OF 12 DICE, IN THE SECOND THROWS OF WHICH 10 DICE WERE LEFT DOWN AND COUNTED *

| | | | Second Throws | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | | Total→ | 1 | 2 | 7 | 24 | 55 | 93 | 111 | 100 | 64 | 31 | 11 | 1 | |
| | 0 | ↓ 1 | 1 | — | — | — | — | — | — | — | — | — | — | — | — |
| | 1 | 1 | — | 1 | — | — | — | — | — | — | — | — | — | — | — |
| | 2 | 7 | — | — | 2 | 5 | — | — | — | — | — | — | — | — | — |
| | 3 | 24 | — | 1 | 3 | 8 | 9 | 3 | — | — | — | — | — | — | — |
| First | 4 | 55 | — | — | 2 | 10 | 18 | 19 | 6 | — | — | — | — | — | — |
| Throws | 5 | 110 | — | — | — | 1 | 24 | 43 | 32 | 10 | — | — | — | — | — |
| | 6 | 93 | — | — | — | — | 4 | 22 | 37 | 24 | 6 | — | — | — | — |
| | 7 | 96 | — | — | — | — | — | 6 | 27 | 39 | 19 | 5 | — | — | — |
| | 8 | 60 | — | — | — | — | — | — | 9 | 17 | 24 | 9 | 1 | — | — |
| | 9 | 42 | — | — | — | — | — | — | — | 10 | 14 | 11 | 7 | — | — |
| | 10 | 10 | — | — | — | — | — | — | — | — | 1 | 6 | 2 | 1 | — |
| | 11 | 1 | — | — | — | — | — | — | — | — | — | — | 1 | — | — |
| | 12 | | — | — | — | — | — | — | — | — | — | — | — | — | — |

* See note to Table 66.

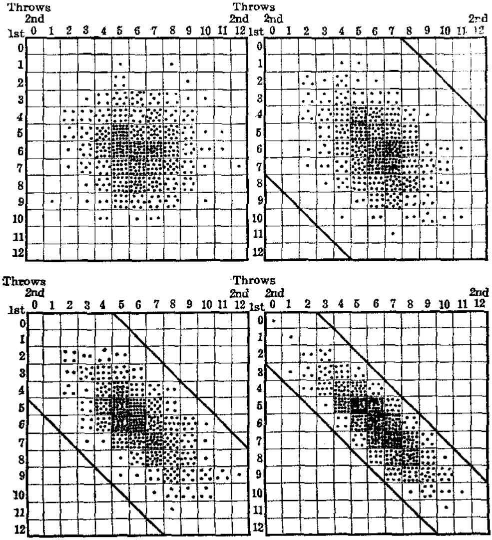
such circumstances often reveal a partial story, are not comparable from time to time and from place to place, and taken alone constitute a weak and uncertain base upon which to establish cause-and-effect relations.

FIGURE 77

Graphic Figures Illustrating Correlation by Means of 500 Pairs of Throws of Dice

Second Throws Independent of First Throws.

Second Throws Dependent on First Throws—5 Dice in Common.




Second Throws Dependent on First Throws—8 Dice in Common.

Second Throws Dependent on First Throws—10 Dice in Common.

## IV. The Measurement of Correlation

Correlation and narrow causation are different. Whether phenomena stand in the relation of "cause and effect" and if so which is cause and which is effect can never be determined statistically.[1] Correlation, or association between them may, however, be determined in this manner. It is quite as possible in two or more series to measure the congruence of change of the corresponding items from a norm or standard as it is to describe in a single series the manner in which the deviations are distributed about a mean. Indeed for both, much the same type of reasoning applies.

### 1. the "sum product" method

In order to understand the measure of correlation most commonly used in statistical analysis, it is necessary very briefly to describe the conditions under which it was developed.

#### *(1) The Assumptions Upon Which the Pearsonian Coefficient of Correlation is Based*

What has come to be known as the Pearsonian coefficient of correlation was conceived by Sir Francis Galton in connection with his work on heredity. In the form in which it is now used it is the creation of Karl Pearson, the English biometrician and statistician. It has since become the tool of biometricians,[2] zoologists,[3] breeders,[4] psychologists,[5] and econ-

[1] *Cf.*, Hooker, "Correlation of the Marriage Rate with Trade," *Journal of the Royal Statistical Society*, Vol. 64, p. 485.

[2] See the journal *Biometrika* and the writings of Sir Francis Galton, Karl Pearson, C. B. Davenport, H. M. Vernon, *et al.*

[3] Among the leading is Harris, J. A., of the Carnegie Institution of Washington, D. C. See his "An Outline of Current Progress in the Theory of Correlation and Contingency," in *American Naturalist*, January, 1916, Vol. L, pp. 53-64.

[4] Davenport, Eugene, *The Principles of Breeding*, Ginn, New York, 1907.

[5] Thorndike, E. L., *Mental and Social Measurements*, New York, 1913; Brown, William, *The Essentials of Mental Measurement*, Cambridge (England), 1911; Whipple, Guy M., *Manual of Mental and Physical Tests*, Baltimore, 1914.

omists.[1] Pearson, in explaining what is meant by correlation, says:

"Two organs in the same individual, or in a connected pair of individuals, are said to be correlated, when a series of the first organ of a definite size being selected, the mean of the sizes of the corresponding second organs is found to be a function of the size of the selected first organ. If the mean is independent of this size, the organs are said to be non-correlated. Correlation is defined mathematically by any constant, or series of constants, which determine the above function."[2]

As Pearson explains, the word "organ" is understood to cover any measurable characteristic of an organism, and the word "size" its quantitative value.

The concepts have been illustrated by Professor Persons as follows:

"Suppose that we are attempting to answer the question, do tall fathers have tall sons? In this case, stature is the 'measurable characteristic' in each of 'a connected pair of individuals.' Suppose the average stature of all adult males is sixty-six inches; suppose we select several thousand fathers whose stature is seventy-two inches or more, six inches above the average for all, and find the mean stature of the sons of this group of tall fathers to be sixty-nine inches, three inches above the average stature of all adult males. If similar results appear consistently for selected fathers and their sons, we may conclude that the stature of sons depends upon the stature of fathers; or, in other words, the stature of sons is a function of the statures of fathers; or, in still other words, the statures of fathers and sons are correlated. We may be able to state a 'law' of the inheritance of stature, or give the stature of sons as a function of the

[1] Hooker, R. H., *op. cit.*; Yule, *Introduction to Theory of Statistics*, London, 1911; Bowley, A. L., *Measurement of Groups and Series*, London, 1903; Elderton, W. Palin, *Frequency Curves and Correlation*, London, 1906 (?); Persons, W. M., "The Correlation of Economic Statistics," *Publications of the American Statistical Association*, Vol. XII, December, 1910, pp. 287-322; Moore, H. L., *Economic Cycles: Their Law and Cause*, New York, 1914; Persons, Warren M., "The Construction of a Business Barometer Based upon Annual Data," in *American Economic Review*, December, 1916, pp. 739-769. See also the notes and references to Chapter XIV.

[2] Pearson, Karl, "Mathematical Contributions to the Theory of Evolution, III. Regression, Heredity, and Panmixia," *Philosophical Transactions of the Royal Society of London*, 1896, A. 187, p. 257.

stature of fathers. It is clear, however, that although tall fathers may, *in general*, have tall sons, an individual tall father may have a short son, or perhaps several sons, some tall, some short. That is, *two concepts are involved; first, the law or function or equation* expressing the relation *on the average*, existing between the two variables involved, and second, the degree with which individual cases adhere to the law.

"To illustrate the first concept, it may be possible to say that for an average deviation in stature of fathers of $n$ inches from the mean for adult males, the stature of the sons of those fathers will deviate in the same direction by $\frac{n}{2}$ inches. This is a law or function, the first concept that we have named.[1] But the statement of the function does not describe the situation completely. How *accurately* does the function describe the situation; how *systematic* is the relationship between statures of fathers and sons; are the exceptional cases few or many? These are different forms of a question which requires a quantitative answer. Such an answer is given by the coefficient of correlation. The coefficient is unity if there is no exception to the law of statures; it is zero if the statures of father and son are independent of each other; it is negative if tall fathers, in general, have short sons; it has a numerical value varying inversely with the degree of divergence (both in number of cases and magnitude) of the individual cases from a linear relationship."[2]

In the language used by Professor Persons, certain expressions appear which were explained in earlier chapters. For instance, "all adult males" (a "population"); "average stature" (a mean or standard); "six inches above the average" (deviations from an average); "mean stature" (an average or norm); "on the average" (an expression indicative of consistency of occurrence—high probability); "how systematic is the relationship between statures of fathers and sons" (an expression indicative of the nature of dispersion). That is, the illustration applies to (1) paired or connected populations or samples; (2) averages characteristic of both; (3) some measurements of the deviations from their respective averages;

[1] Note omitted.
[2] Persons, W. M., "Indices of Business Conditions," *Review of Economic Statistics*, Cambridge, Mass., January, 1919, p. 131.

(4) systematic and regular distribution of the deviations from their averages; and (5) a measurement of the congruence of change in the corresponding deviations in both samples from their respective averages. Now, it is apparent that by the use of some of these statistical devices, frequency and other distributions are measured and compared. Only two new ideas are introduced—(1) connected or related series, and (2) the measurement of concurrent deviations.

The Pearsonian coefficient of correlation rests upon two assumptions. The first is that a large number of independent causes are operating in each of the series correlated so as to produce normal or probability distributions. Such causes are at work in determining the successive results secured by Darbishire in throwing his twelve dice. They undoubtedly are also operating to produce the heights of both fathers and sons in Professor Persons' illustration. Such series, as we have learned, can be summarized by the use of averages and by measures and coefficients of dispersion.

The second assumption latent in the Pearsonian coefficient is that the forces so operating are not independent of each other —in the random sense—but that they are related in a causal way. This is evidently the case in the second throws of dice wherein some of the "effects" are determined by the conditions in the first throws—chance, however, having fully operated to produce the result. It is also true in the case of the heights of sons if they are correlated with those of their fathers.

To count in the second dice throws part of the results secured in the first throws does not have the effect of producing distributions any less normal in the second throws. Chance operates just the same. The only thing which is done in the illustration is to transfer to the chance distribution in the second throws some of the chance results in the first throws. Any throw is as much governed by chance as any other throw. Accordingly, such a transfer is legitimate. Similarly, the heights of a large number of fathers tend to conform

to the normal probability curve. Such a condition may also be expected of those of their sons. If the forces producing these results are not independent of each other, then it is said that the heights of sons are correlated with those of their fathers.

Upon the bases of these two assumptions, Pearson constructed the formula for his coefficient. His own words in respect to the organs correlated are as follows:

> The assumptions are: first, "that the sizes of this complex of organs are determined by a great variety of *independent* contributory causes, for example, magnitudes of other organs not in the complex, variations in environment, climate, nourishment, physical training, and innumerable other causes, which cannot be individually observed or their effects measured"; second, "that the variations in intensity of the contributory causes are small as compared with their absolute intensity, and that these variations follow the normal law of distribution."[1]

### (2) *The Pearsonian Coefficient of Correlation Formula*

The Pearsonian coefficient of correlation formula[2] is $r = \frac{\Sigma xy}{n \sigma_1 \sigma_2}$ where

$r$ = the coefficient of correlation
$xy$ = the product of a concurrent pair of deviations
$\Sigma$ = the process of summation
$\sigma_1$ = the standard deviation, *S.D.*, of one ($X$) series
$\sigma_2$ = the standard deviation, *S.D.*, of the other ($Y$) series
$n$ = total number of pairs of items

This formula gives values ranging from $-1$ through 0 to $+1$.[3]

When $\Sigma xy$ is positive, correlation is positive; when it is negative, correlation is negative. Positive eorrelation may re-

[1] Pearson, *op. cit.*, p. 262.
[2] For the method by which this formula is derived, see Yule, G. Udny, *Introduction to the Theory of Statistics*, Griffin, London, 1911, pp. 168-174.
[3] For proof of this, see Bowley, A. L., *Elements of Statistics*, 4th Ed., King, London, 1920, p. 354.

sult from positive items (that is, items larger than the mean) in one ($X$) series being associated with positive items (that is, items larger than the mean) in the other ($Y$) series, or from negative items (those smaller than the mean) in one ($X$) series being associated with negative items (those smaller than the mean) in the other ($Y$) series. Negative correlation results from positive values (those larger than the mean) in one ($X$) series being associated with negative values (items smaller than the mean) in the other ($Y$) series, or vice versa.

When positive and negative deviations in the two series are indifferently associated, correlation tends to be zero, reaching this limit when the negative products exactly counterbalance those which are positive.

It should be noticed that the sum of the products of the *deviations*—$\Sigma xy$—*is a function both of the amount and sign* of the deviations. Moreover, since the deviations are taken from the respective means of the series and these may differ not only in size but also in the unit of measurement, some divisor is necessary in order to reduce them to the same denomination. The standard deviation in each case is the appropriate factor here as it is in the measurement of relative dispersion. But since the deviations are multiplied together, the suitable divisor is the product of the standard deviations. Correlation coefficients, however, are compared for series with different numbers of pairs of items. Accordingly, $n$ is inserted in the denominator, thus giving an average value independent of the number. Accordingly, the correlation coefficient—$r$—of two sets of values, each expressed in standard deviations as units, is the arithmetic average of the products of deviations of corresponding values from their respective means.

"Hence $r$ is a quantity which depends on all the observations, is zero when independence is complete and *Mean* $xy = 0$, is independent of the units in which $X$ and $Y$ are measured, increases whenever a positive $x_t$ is found with a positive $y_t$ or a negative $x_t$ with a negative $y_t$, but only reaches the value $+1$ (which it can never exceed) when $x$ and $y$ are connected rigidly by the equation $y = x \times$ constant. If

positive $x$'s are found with negative $y$'s and *vice versa*, $r$ varies from o to — 1.

"$r$ is therefore a sensitive measurement of the amount of correlation."[1]

Now it is apparent from Table 66 that the two series, first throws ($Y$) and second throws ($X$), are not correlated. That is, neither high nor low values in ($Y$) are associated with high or low values or vice versa in ($X$). With essentially the same values in the second throws varying values are found in the first throws. Similarly, with essentially the same values in the first throws different values are found in the second throws. Relations are different in Table 69. In this case, as the values in the first throws increase so do those in the second throws. That is, the two series are positively correlated. If with increases in the first were found decreases in the second, or vice versa, then the two series would be negatively correlated. If the association were not greater than that secured from random selection—as in Table 66—correlation would be small, the coefficient approaching zero.

While such frequency tables as 68 and 69 indicate correlation, they do not measure it. The fact of correlation is generally evident from the nature of the distribution of the frequencies in the lines and in the columns. If the area of concentration extends from the upper left to the lower right corners of the frequency surface, then correlation is positive; if from the upper right to the lower left, it is negative.[2] If neither arrangement is apparent, as in Table 66, correlation is small and the type in doubt.

Moreover, if correlation is present, the arithmetic means and the medians of the rows and of the columns form a more

[1] Bowley, A. L., *Elements of Statistics*, 4th Ed., King, London, 1920, pp. 354-355.

[2] The nature of correlation, that is, whether positive or negative, as indicated by the direction which the concentration takes, is obviously determined by the ways in which the scales on the respective axes are written.

or less regular progression.[1] By this test the first and second throws of dice, as shown in Table 66, are not correlated. The medians in the rows and columns are constant at about 5-6. On the other hand, in the series in Table 69—which are known to be highly correlated—the progression of the medians of the rows and columns is strikingly regular. In terms of averages, large values in one series are associated with large values in the other series.

In general, if the average values for the detail in the rows and columns are linear—best described by straight lines—then the Pearsonian coefficient is a suitable measure for measuring correlation. The coefficient may be computed both from grouped and ungrouped data. While the methods are somewhat different, the principles are identical.

### (3) *The Calculation of the Pearsonian Coefficient of Correlation*

#### a. In Ungrouped Series

In an address on *Concentration of Power Supply,* Mr. Samuel Insull, President of the Commonwealth Edison Company, Chicago, said in relation to statistics there considered: "The income per kilowatt hour goes down pretty steadily, the output per capita goes up pretty steadily, the load factor improves as selling price is lowered, and the output per capita goes up as the selling price is lowered."[2] These conclusions were based upon a consideration of the United States Census figures for 1912 on the generation of electrical energy giving the capacity load factor,[3] output per capita, and income per kilowatt hour by states. It is the correlation of the load fac-

[1] See Figure 78.

[2] Address before the Finance Forum of the Young Men's Christian Association, New York, 1914, privately printed, p. 26.

[3] Ratio of average load to capacity in this case, p. 26.

TABLE 70

TABLE SHOWING BY STATES THE CAPACITY LOAD FACTOR AND THE INCOME PER KILOWATT HOUR IN THE GENERATION OF ELECTRICAL ENERGY

| STATE | CAPACITY LOAD FACTOR % $X$ | DEVIATIONS FROM AVERAGE LOAD FACTOR $x$ | DEVIATIONS SQUARED $x^2$ | INCOME PER K.W.H. (in cents) $Y$ | DEVIATIONS FROM AVERAGE INCOME PER K.W.H. $y$ | DEVIATIONS SQUARED $y^2$ | PRODUCT OF DEVIATIONS ($x$'s) and ($y$'s) |
|---|---|---|---|---|---|---|---|
| Total .... | av. 21.4 | | 4144.61 | av. 3.45 | | 177.2011 | — 444.735 |
| Alabama ... | 22.7 | + 1.3 | 1.69 | 2.49 | — .96 | .9216 | — 1.248 |
| Arizona .... | 25.4 | + 4.0 | 16.00 | 3.56 | — .11 | .0121 | — .440 |
| Arkansas ... | 12.4 | — 9.0 | 81.00 | 5.45 | + 2.00 | 4.0000 | — 18.000 |
| California .. | 33.9 | + 12.5 | 156.25 | 1.59 | — 1.86 | 3.4596 | — 23.250 |
| Colorado ... | 25.3 | + 3.9 | 15.21 | 2.89 | — .56 | .3136 | — 2.184 |
| Conn. ...... | 19.2 | — 2.2 | 4.84 | 4.10 | + .65 | .4225 | — 1.430 |
| Florida .... | 12.5 | — 8.9 | 79.21 | 5.11 | + 1.66 | 2.7556 | — 14.774 |
| Georgia .... | 17.8 | — 3.6 | 12.96 | 2.01 | — 1.44 | 2.0736 | + 5.184 |
| Idaho ...... | 37.0 | + 15.6 | 243.36 | 1.37 | — 2.08 | 4.3264 | — 32.448 |
| Illinois ..... | 29.3 | + 7.9 | 62.41 | 2.52 | — .93 | .8649 | — 7.347 |
| Indiana .... | 19.9 | — 1.5 | 2.25 | 3.26 | — .19 | .0361 | + .285 |
| Iowa ...... | 14.4 | — 7.0 | 49.00 | 6.45 | + 3.00 | 9.0000 | — 21.000 |
| Kansas ..... | 22.0 | + .6 | .36 | 2.19 | — 1.26 | 1.5876 | — .756 |
| Kentucky .. | 15.9 | — 5.5 | 30.25 | 3.64 | + .19 | .0361 | — 1.045 |
| Louisiana ... | 10.9 | — 10.5 | 110.25 | 12.25 | + 8.80 | 77.4400 | — 92.400 |
| Maine ..... | 22.7 | + 1.3 | 1.69 | 1.74 | — 1.71 | 2.9241 | — 2.223 |
| Maryland .. | 5.0 | — 16.4 | 268.96 | 1.37 | — 2.08 | 4.3264 | + 34.112 |
| Mass. ...... | 17.5 | — 3.9 | 15.21 | 4.17 | + .72 | .5184 | — 2.808 |
| Mich. ...... | 23.2 | + 1.8 | 3.24 | 2.19 | — 1.26 | 1.5876 | — 2.268 |
| Minn. ...... | 22.7 | + 1.3 | 1.69 | 3.72 | + .27 | .0729 | + .351 |
| Miss. ...... | 14.6 | — 6.8 | 46.24 | 4.02 | + .57 | .3249 | — 3.876 |
| Missouri ... | 21.7 | + .3 | .09 | 4.18 | + .73 | .5329 | + .219 |
| Montana ... | 58.0 | + 36.6 | 1339.56 | 1.05 | — 2.40 | 5.7600 | — 87.840 |
| Nebraska ... | 18.6 | — 2.8 | 7.84 | 4.98 | + 1.53 | 2.3409 | — 4.284 |
| Nevada .... | 48.6 | + 27.2 | 739.84 | 1.38 | — 2.07 | 4.2849 | — 56.304 |
| New Ham... | 25.0 | + 3.6 | 12.96 | 1.84 | — 1.61 | 2.5921 | — 5.796 |
| New Jersey. | 24.4 | + 3.0 | 9.00 | 2.85 | — .60 | .3600 | — 1.800 |
| New Mex... | 12.9 | — 8.5 | 72.25 | 5.50 | + 2.05 | 4.2025 | — 17.425 |
| New York.. | 32.1 | + 10.7 | 114.49 | 2.63 | — .82 | .6724 | — 8.774 |

TABLE 70 (*Continued*)

| State | Capacity Load Factor % $X$ | Deviations from Average Load Factor $x$ | Deviations Squared $x^2$ | Income per K.W.H. (in cents) $Y$ | Deviations from Average Income per K.W.H. $y$ | Deviations Squared $y^2$ | Product of Deviations ($x$'s) and ($y$'s) |
|---|---|---|---|---|---|---|---|
| N. Car...... | 18.7 | — 2.7 | 7.29 | 1.90 | — 1.55 | 2.4025 | + 4.185 |
| N. Dakota.. | 12.9 | — 8.5 | 72.25 | 7.01 | + 3.56 | 12.6736 | — 30.260 |
| Ohio ....... | 18.6 | — 2.8 | 7.84 | 2.99 | — .56 | .3136 | + 1.568 |
| Oklahoma .. | 19.7 | — 1.7 | 2.89 | 4.54 | + 1.09 | 1.1881 | — 1.836 |
| Oregon ..... | 20.7 | — .7 | .49 | 2.39 | — 1.06 | 1.1236 | + .742 |
| Penn. ...... | 15.7 | — 5.7 | 32.49 | 4.14 | + .69 | .4761 | — 3.933 |
| Rhode Island | 18.4 | — 3.0 | 9.00 | 3.71 | + .26 | .0676 | — .780 |
| S. Carolina.. | 30.7 | + 9.3 | 86.49 | 1.24 | — 2.21 | 4.8841 | — 20.553 |
| S. Dakota... | 14.0 | — 7.4 | 54.76 | 4.58 | + 1.13 | 1.2769 | — 8.362 |
| Tenn. ...... | 17.4 | — 4.0 | 16.00 | 3.24 | — .21 | .0441 | + .840 |
| Texas ...... | 27.6 | + 6.2 | 38.44 | 3.38 | — .07 | .0049 | — .434 |
| Utah ...... | 26.0 | + 4.6 | 21.16 | 1.75 | — 1.70 | 2.8900 | — 7.820 |
| Vermont ... | 21.9 | + .5 | .25 | 2.07 | — 1.38 | 1.9044 | — .690 |
| Virginia .... | 8.1 | — 13.3 | 176.89 | 2.65 | — .80 | .6400 | + 10.640 |
| Wash. ...... | 14.2 | — 7.2 | 51.84 | 4.33 | + .88 | .7744 | — 6.336 |
| West Va.... | 16.1 | — 5.3 | 28.09 | 2.60 | — .85 | .7225 | + 4.505 |
| Wisconsin .. | 24.9 | + 3.5 | 12.25 | 2.92 | — .53 | .2809 | — 1.855 |
| Wyoming ... | 16.1 | — 5.3 | 28.09 | 6.24 | + 2.79 | 7.7841 | — 14.787 |

tor and the income per K.W.H. which is measured in Table 70 and the accompanying computations.[1]

In this case the "capacity load factor" constitutes one ($X$) series, and the "income per K.W.H." the other ($Y$) series. The steps in calculating the coefficient of correlation are as follows:

1. Determine the arithmetic mean in each of the series.
2. Calculate the deviations (differences) of each of the items in the series from their respective arithmetic means.

[1] These figures are inadequate for a satisfactory study of this character. They will, however, serve to illustrate the manner in which similar data may be compared.

(The deviations of the items in the $X$ series are given in the column marked $x$; for those in the $Y$ series, in the column marked $y$).

3. *Square the deviations for each of the series.* See columns marked ($x^2$) and ($y^2$).

4. Multiply together the corresponding deviations for the $X$ and the $Y$ series (that is, the amounts in columns $x$ and $y$).

5. Algebraically sum or total the products obtained in 4.

The total secured from step five gives the numerator—$\Sigma\, xy$—of the coefficient. But the standard deviation in each series is also required. This is determined by using the formula,[1] $\sqrt{\frac{\Sigma\, d^2}{n}}$. In the illustration in Table 70 the $d$'s in the $X$ series are called $x$'s; those in the $Y$ series, $y$'s. Accordingly, the formula for the $X$ series is $\sqrt{\frac{\Sigma\, x^2}{n}}$; for the $Y$ series, $\sqrt{\frac{\Sigma\, y^2}{n}}$.

Each of the amounts required for the coefficients are now available *except the $n$ of the denominator.* $n$ means the number of pairs of values—in this case 47, since there are 47 states for which data are available.

The standard deviation of the $X$ series is $\sqrt{\frac{\Sigma\, x^2}{n}}$ or $\sqrt{\frac{4144.61}{47}} = 9.39$; that of the $Y$ series, $\sqrt{\frac{\Sigma\, y^2}{n}}$ or $\sqrt{\frac{177.2011}{47}} = 1.95$. Inserting these and the other appropriate values in the formula, $r = \frac{\Sigma\, xy}{n\, \sigma_1\, \sigma_2}$, gives $r = \frac{-444.735}{47 \times 9.39 \times 1.95}$ $= -0.517$. That is, the two series are negatively correlated.

The quantity[2], $-0.517$, is a measure of the congruence of

[1] See p. 349.

[2] For a discussion of the "significance" of this coefficient, see pp. 428-430.

change in the deviations of the items in the two series. It is the mean of the products of the deviations—measured from the averages of the series—expressed in units of standard deviations. The negative sign (—) indicates that on the average, positive and negative deviations, or vice versa, are associated. The decimal (0.517) shows the degree of such association. If an increase in one series were associated with a proportional decrease in the other series, or vice versa, the ratio would be — 1.

### b. In Grouped Series

The ungrouped series in Table 70 might be tabulated in double frequency form similar to the tables showing throws of dice. If this were done, provision would be made in the stub (or the caption) for the load factor per cents, and in the caption (or the stub) for the K.W.H. amounts. The states would then be tallied in columns and rows according to the unit classes in stub and caption.

In order to show the method of calculating Pearson's $r$ for grouped data, rental payments made by retail clothing stores are used. The question upon which information is desired is as follows: In what manner and to what degree, if any, in retail clothing stores are the amounts of rent paid in units of sales correlated with rental payments in units of floor space? A sample of 150 stores is used.

The data for the different stores might be arranged in the form shown in Table 70. They would then appear as follows:

| Store | Rent per $100 of Sales | Rent per 100 sq. ft. of Floor Space |
|---|---|---|
| 1 | $ .75 | $30.00 |
| 2 | 1.25 | 32.00 |
| 3 | .92 | 34.00 |
| 4 | .87 | 36.00 |
| etc. | etc. | etc. |

In this case, however, the form of arrangement selected is the double frequency table 71.

TABLE

TABLE SHOWING THE METHOD OF CALCULATING THE CORRELATION
THE TRUE

| Amounts of Rent / Y Series—Per $100 of Sales | X Series Per 100 Square Feet of Floor Space: Under $5 | $5–10 | 10–15 | 15–20 | 20–25 | 25–30 | 30–35 | 35–40 | 40–45 | 45–50 | 50–55 | 55–60 | 60–65 | 65–70 | 70–75 | 75 & over | Frequencies (Total) $f$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $3.00 & over | — | — | — | — | — | 1 | 1 | 1 | 1 | 2 | — | 2 | — | — | — | 8 | 16 |
| 2.80–$3.00 | — | — | — | — | — | — | — | 1 | 1 | 1 | — | — | 1 | 1 | — | — | 5 |
| 2.60– 2.80 | — | — | — | — | 1 | — | — | — | — | — | — | 1 | — | — | — | 2 | 4 |
| 2.40– 2.60 | — | — | — | — | — | — | 1 | — | — | 1 | — | — | — | — | — | 1 | 3 |
| 2.20– 2.40 | — | — | — | — | — | 1 | 2 | — | 2 | — | 1 | 1 | 3 | — | — | — | 10 |
| 2.00– 2.20 | — | — | — | 1 | — | — | 3 | 1 | — | 2 | — | — | — | — | 1 | 3 | 11 |
| 1.80– 2.00 | — | — | — | — | 2 | 1 | 3 | 3 | 2 | 3 | — | — | 1 | — | — | — | 15 |
| 1.60– 1.80 | — | — | — | — | 3 | 1 | 6 | 1 | 5 | — | 1 | — | — | — | — | — | 17 |
| 1.40– 1.60 | 1 | — | 2 | 2 | 1 | 2 | 2 | 1 | — | — | 2 | 1 | — | — | 1 | — | 15 |
| 1.20– 1.40 | — | — | 2 | 1 | 3 | 4 | 3 | 1 | 4 | — | 1 | 2 | — | — | — | — | 21 |
| 1.00– 1.20 | — | — | 1 | 3 | 5 | 2 | 3 | 2 | 1 | — | — | — | — | — | — | — | 17 |
| .80– 1.00 | — | 1 | — | 2 | 3 | 4 | 1 | — | — | — | — | — | — | — | — | — | 11 |
| .60– .80 | — | 1 | — | 1 | 1 | 1 | 1 | — | — | — | — | — | — | — | — | — | 5 |
| | | | | | | | | | | | | | | | | | 150 |

| X Series | Frequencies (Total) $f$ | Deviations from Arith. Mean $d$ | Deviations Squared $d^2$ | Deviations Squared Times Frequencies $fd^2$ |
|---|---|---|---|---|
| Under $5 | 1 | − 36.1 | 1,303 | $ 1,303 |
| $5–10 | 2 | − 31.1 | 967 | 1,934 |
| 10–15 | 5 | − 26.1 | 681 | 3,405 |
| 15–20 | 10 | − 21.1 | 445 | 4,450 |
| 20–25 | 19 | − 16.1 | 259 | 4,921 |
| 25–30 | 17 | − 11.1 | 123 | 2,091 |
| 30–35 | 26 | − 6.1 | 37 | 962 |
| 35–40 | 11 | − 1.1 | 1 | 11 |
| 40–45 | 16 | + 3.9 | 15 | 240 |
| 45–50 | 9 | + 8.9 | 79 | 711 |
| 50–55 | 5 | + 13.9 | 193 | 965 |
| 55–60 | 7 | + 18.9 | 357 | 2,499 |
| 60–65 | 5 | + 23.9 | 571 | 2,855 |
| 65–70 | 1 | + 28.9 | 835 | 835 |
| 70–75 | 2 | + 33.9 | 1,149 | 2,298 |
| 75 & over | 14 | + 38.9 | 1,513 | 21,182 |
| | 150 | | | $50,662 |

Arith. Mean = $38.6 : X series

$$\text{S.D. or } \sigma_1 = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{\$50{,}662}{150}}$$
$$= \$18.4$$

71

Coefficient for Grouped Series, Deviations Being Taken from Arithmetic Mean

| Deviations from Arith. Mean $d$ | Deviations Squared $d^2$ | Deviations Squared Times Frequencies $fd^2$ | Products of the Respective Deviations in the Two Series $(xy)$ |
|---|---|---|---|
| + 1.31 | 1.72 | $27.52 | + 461.64 |
| + 1.11 | 1.23 | 6.15 | + 71.60 |
| + .91 | .83 | 3.32 | + 73.35 |
| + .71 | .50 | 1.50 | + 29.61 |
| + .51 | .26 | 2.60 | + 45.39 |
| + .31 | .10 | 1.10 | + 39.65 |
| + .11 | .01 | .15 | — .83 |
| — .09 | .01 | .17 | + 5.73 |
| — .29 | .08 | 1.20 | + 29.44 |
| — .49 | .24 | 5.04 | + 57.87 |
| — .69 | .48 | 8.16 | + 144.00 |
| — .89 | .79 | 8.69 | + 153.17 |
| — 1.09 | 1.19 | 5.95 | + 93.20 |
| | | *Total* $71.55 | +1203.82 |

Arith. Mean = $1.79: $Y$ series $r = \dfrac{1203.82}{150 \times .69 \times 18.4}$

S.D. or $\sigma_2 = \sqrt{\dfrac{\Sigma d^2}{n}} = \sqrt{\dfrac{\$71.55}{150}}$

$= \$.69$ $\qquad = \dfrac{+\,1203.82}{1904.40} = +\,.63$

P.E. = $\pm\,.033$

$r = +\,.63 \pm .033$

The arrangement of the data across the surface of Table 71 indicates plainly the *fact* of positive correlation.[1] But what is the degree of correlation? Pearson's $r$ gives this in precise form.

*The steps to be taken in securing $r$ are as follows:*

1. Total the frequencies in the rows—that is, the numbers of stores paying different amounts of rent per 100 square

[1] Notice the method of writing the stub classes. Positive correlation in this case is indicated by a different alignment from that in Table 68, for instance.

feet of floor space. The totals are 16, 5, 4, 3, 10, 11, 15, 17, 15, 21, 17, 11, 5. Total, 150.

2. Total the frequencies in the columns—that is, the numbers of stores paying different amounts of rent per $100 of sales. The totals are 1, 2, 5, 10, 19, 17, 26, 11, 16, 9, 5, 7, 5, 1, 2, 14. Total, 150.

3. For each of these frequency distributions calculate the mean—use the center of the group for precise items—and the standard deviation. The methods by which these computations are made have already been explained. (In order to get *S.D.* in each case, each deviation must be multiplied by the number of corresponding frequencies.)

4. Calculate the products of the corresponding deviations from the means in the two series. The items in the table deviate from the averages of both series. For instance, the 10 instances in which rent as a per cent of sales is 2.30 deviate from the average for the entire series, 1.79, by +.51. At the same time, they also deviate from the average of the series showing the amount of rent paid per 100 square feet of floor space. The average in this case is $38.6. One of them deviates − 11.1; 2 of them − 6.1; 2, + 3.9; 1, + 13.9; 1, + 18.9; and 3, +23.9. That is, to get the products of the deviations of the ten items from the averages in the series it is necessary to make the following computations:

$$\left.\begin{array}{l} 1 \times -11.1 \\ 2 \times -\ \ 6.1 \\ 2 \times +\ \ 3.9 \\ 1 \times +13.9 \\ 1 \times +18.9 \\ 3 \times +23.9 \end{array}\right\} \times +.51 = 45.39$$

The other amounts in the column, $(xy)$, are secured in similar manner.

5. Algebraically sum or total the products secured in 4 above. See the total of column $(xy)$.

If the values secured by the above processes are inserted

in the formula, $r = \frac{\Sigma xy}{n\,\sigma_1\,\sigma_2}$, the result is as follows:

$$r = \frac{+\,1203.82}{150 \times .69 \times 18.4} = +\,.63$$

That is, correlation is positive—the + sign indicating this fact.

But it is sometimes advantageous to compute $r$ for grouped series by assuming the arithmetic means, and later by correcting in each of the steps the errors due to the assumption. The manner in which this is done is illustrated in Table 72 by using the data given in Table 71.[1]

The notation used is as follows:

$f$ = frequencies in the $X$ and in the $Y$ series
$x$ = deviations in steps (groups) in the $X$ series
$y$ = deviations in steps (groups) in the $Y$ series
$\Sigma$ = process of summation
$d_x$ = average error of deviations in the $X$ series
$d_y$ = average error of deviations in the $Y$ series

The steps in computing $r$ by this method are as follows:

1. Total the frequencies of the lines and of the columns. (See column $f$ and line $f$).

2. Choose an average (group) in the $X$ and in the $Y$ series, respectively. Draw lines at right angles across the table enclosing the frequencies in these groups.

3. Indicate the group deviations above and below the assumed average (group). See column $y$ for the deviations for the $Y$ series; and line $x$ for the deviations for the $X$ series.

4. Multiply the frequencies in the two series by their respective group deviations. See column $fy$ for the $Y$ series, and line $fx$ for the $X$ series.

5. Square the group deviations in the two series and multiply by their respective frequencies. See column $fy^2$ for the $Y$ series, and $fx^2$ for the $X$ series.

6. Compute the amount and nature (plus (+) or minus

[1] The method of computing the arithmetic mean from an assumed average is shown in Table 37. Similarly, the method of computing the standard deviation when an assumed mean is used is illustrated in Table 63.

TABLE

TABLE SHOWING THE METHOD OF CALCULATING THE CORRELATION
BEING TAKEN FROM AN

| OUNTS OF RENT | X SERIES PER 100 SQUARE FEET OF FLOOR SPACE Under $5 | $5–10 | 10–15 | 15–20 | 20–25 | 25–30 | 30–35 | 35–40 | 40–45 | 45–50 | 50–55 | 55–60 | 60–65 | 65–70 | 70–75 | 75 & over | f | y | fy | fy² |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | | | | |
| | | | | | | | | | | | | | | | | | | | Total 177 | |
| 3.00 & over | — | — | — | — | — | 1 | 1 | 1 | 1 | 2 | — | 2 | — | — | — | 8 | 16 | 6 | 96 | 576 |
| 2.80–$3.00 | — | — | — | — | — | — | — | 1 | 1 | 1 | — | — | 1 | 1 | — | — | 5 | 5 | 25 | 125 |
| 2.60– 2.80 | — | — | — | — | 1 | — | — | — | — | — | — | 1 | — | — | — | 2 | 4 | 4 | 16 | 64 |
| 2.40– 2.60 | — | — | — | — | — | — | 1 | — | — | 1 | — | — | — | — | — | 1 | 3 | 3 | 9 | 27 |
| 2.20– 2.40 | — | — | — | — | — | 1 | 2 | — | 2 | — | 1 | 1 | 3 | — | — | — | 10 | 2 | 20 | 40 |
| 2.00– 2.20 | — | — | — | 1 | — | — | 3 | 1 | — | 2 | — | — | — | — | 1 | 3 | 11 | 1 | 11 | 11 |
| 1.80– 2.00 | — | — | — | — | 2 | 1 | 3 | 3 | 2 | 3 | — | — | 1 | — | — | — | 15 | 0 | | |
| 1.60– 1.80 | — | — | — | — | 3 | 1 | 6 | 1 | 5 | — | 1 | — | — | — | — | — | 17 | 1 | 17 | 17 |
| 1.40– 1.60 | 1 | — | 2 | 2 | 1 | 2 | 2 | 1 | — | — | 2 | 1 | — | — | 1 | — | 15 | 2 | 30 | 60 |
| 1.20– 1.40 | — | — | 2 | 1 | 3 | 4 | 3 | 1 | 4 | — | 1 | 2 | — | — | — | — | 21 | 3 | 63 | 189 |
| 1.00– 1.20 | — | — | 1 | 3 | 5 | 2 | 3 | 2 | 1 | — | — | — | — | — | — | — | 17 | 4 | 68 | 272 |
| .80– 1.00 | — | 1 | — | 2 | 3 | 4 | 1 | — | — | — | — | — | — | — | — | — | 11 | 5 | 55 | 275 |
| *.60– .80* | — | *1* | — | *1* | *1* | *1* | *1* | — | — | — | — | — | — | — | — | — | *5* | *6* | *30* | *180* |
| | | | | | | | | | | | | | | | | | | | 263 Total | 1836 Total |
| f | 1 | 2 | 5 | 10 | 19 | 17 | 26 | 11 | 16 | 9 | 5 | 7 | 5 | 1 | 2 | 14 | 150 150 | | | |
| x | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | | | | |
| fx Total 121 | 6 | 10 | 20 | 30 | 38 | 17 | | 11 | 32 | 27 | 20 | 35 | 30 | 7 | 16 | 126 | Total 304 | | | |
| fx² | 36 | 50 | 80 | 90 | 76 | 17 | | 11 | 64 | 81 | 80 | 175 | 180 | 49 | 128 | 1134 | Total 2251 | | | |

(—)) of the deviations. This is done by multiplying the respective frequencies by the amount of group deviations in the Y series. See columns $\left(\frac{\Sigma x \text{ (gross)}}{- \text{ and } +}\right)$.

7. From the minus (—) and plus (+) entries found in 6, compute the net deviations. See column $\left(\frac{\Sigma x \text{ (net)}}{- \text{ and } +}\right)$.

8. Multiply the net deviations found in 7 by the group deviations (see column $y$) in the Y series (see column $\Sigma xy$). In doing this it is necessary carefully to observe the signs of the products.

2

ʾOEFFICIENT FOR GROUPED SERIES, THE DEVIATIONS
ˌSSUMED ARITHMETIC MEAN

| Deviations | | | | | |
|---|---|---|---|---|---|
| Σx (gross) | | Σx (net) | | Σxy | |
| − | + | − | + | − | + |
| 1 | 91 | | 90 | | 540 |
| | 19 | | 19 | | 95 |
| 2 | 23 | | 21 | | 84 |
| | 12 | | 12 | | 36 |
| 1 | 31 | | 30 | | 60 |
| 3 | 42 | | 39 | | 39 |
| | | | | | |
| 7 | 15 | | 8 | 8 | |
| 24 | 22 | 2 | | | 4 |
| 21 | 23 | | 2 | 6 | |
| 25 | 4 | 21 | | | 84 |
| 21 | | 21 | | | 105 |
| 11 | | 11 | | | 60 |
| | | | | 14 Total | 1113 Total |

| | | |
|---|---|---|
| $\Sigma fx$ pos. $= 304$ | $\Sigma fy$ pos. $= 177$ | $\Sigma xy$ pos. $= 1113$ |
| $\Sigma fx$ neg. $= 121$ | $\Sigma fy$ neg. $= 263$ | |
| $\Sigma fx = 183$ | $\Sigma fy = -86$ | |
| $\Sigma fx^2 = 2251$ | $\Sigma fy^2 = 1836$ | $\Sigma xy$ neg. $= -14$ |
| $d_x = \frac{\Sigma fx}{N} = \frac{183}{150} = 1.2$ <br> $d_x^2 = 1.4$ | $d_y = \frac{\Sigma fy}{N} = \frac{-86}{150} = -.6$ <br> $d_y^2 = .4$ | |
| $\sigma_x^2 = \frac{\Sigma fx^2}{N} - d_x^2$ <br> $= \frac{2251}{150} - 1.4$ <br> $= 15.0 - 1.4 = 13.6$ <br> $= \sqrt{13.6} = 3.7$ | $\sigma_y^2 = \frac{\Sigma fy^2}{N} - d_y^2$ <br> $= \frac{1836}{150} - .4$ <br> $= 12.2 - 4 = 11.8$ <br> $= \sqrt{11.8} = 3.4$ | $\Sigma xy = 1099$ <br> $N = 150$ |

$$r = \frac{\frac{\Sigma xy}{N} - d_x \times d}{\sigma_y \times \sigma_y}$$

$$= \frac{\frac{1099}{150} - (1.2 \times -.6)}{3.7 \times 3.4}$$

$$= \frac{7.3 + .72}{3.7 \times 3.4} = \frac{8.02}{12.58}$$

$$r = +.64$$

$$\text{P. E.} = .6745 \frac{1 - r^2}{\sqrt{n}} = \pm .033$$

$$r = +.64 \pm .033$$

The four quadrants of the correlation table relative to the means are as follows:

| | |
|---|---|
| x = — <br> y = + | x = + <br> y = + |
| x = — <br> y = — | x = + <br> y = — |

Accordingly, the signs in the column $\Sigma\, xy$ are determined by these relations.

The foregoing computations give the deviations from the assumed means and the data based upon them for computing the standard deviations in the two series. But since the positive and negative deviations in the two series do not balance (see the totals in column $fy$, and in line $fx$) the assumed are not the correct averages. The deviations in the $X$ series are too large, and those in the $Y$ series too small. Accordingly, corrections must be made for them. This is done in the blocks at the right of the table. The average error in the deviations in the $X$ series ($d_x$) and the corresponding error in the $Y$ series ($d_y$) must be squared, and subtracted from the average of the respective squared deviations in order to obtain the true standard deviations. The product of these average errors must then be subtracted from the average of the products, $\left(\frac{\Sigma\, xy}{N}\right)$, in order to get the true sum of the products.

These various adjustments are carried out in the computations at the right of the table. While the deviations are taken in groups so also are the $S.\ D.$'s and the $xy$ products. Accordingly, this fact may be ignored in the final result.

The coefficient of correlation between rental payments in units of sales and rental payments per 100 sq. ft. of floor space for the 150 retail clothing stores is as follows:

$$r = \frac{\frac{+\,1099}{150} - (1.2 \times -.6)}{3.7 \quad \times \quad 3.4}$$

$$= \frac{+\,8.02}{12.58}$$

$$= +\,.64^{1}$$

[1] This amount differs slightly from that secured in Table 71 because of adjustments of decimal amounts.

### (4) *Regression Lines and Coefficients of Regression*

But correlation between the series in Tables 71 and 72 is not perfect. The means—best values—of the columns and rows are not identical as they would be if perfect correlation existed.[1] In Figure 78 the means of the rows are indicated by crosses (x x) for different values in the *Y* series. Similarly, the means of the columns are indicated by circles (o o) for different values in the *X* series. If perfect positive correlation obtained, the means would fall on a single straight line. As it is, two lines are necessary to show the relations, both the crosses (x x) and the circles (o o) being essentially linear in their arrangement. The best indication of the directions which they take are straight lines so drawn that the sums of the squares of the differences, measured parallel to the *Y* axis, of the several points from the lines are a minimum. These are the "best fitting lines" under the least-square assumption.[2]

If the respective deviations in each series, *X* and *Y*, from their means were expressed in units of standard deviations—that is, if each of them were divided by the standard deviation of the series to which it belongs—and plotted to a scale of standard deviations, plus (+) and minus (—), the slope of a straight line, best describing the plotted points, would be the correlation coefficient, *r*.

The best fitting line of the means of the rows is AB, and of the means of the columns, CD. These are the so-called "regression lines,"[3] their slopes being expressed in terms of

[1] If, in the case of the dice throws, the second throws were taken to be *equivalent to the first throws*, then the means of the columns would be the same as the means of the rows. See Table 69 in which ten of the dice in the first throws are counted in the second throws.

[2] The sum of the squares of the deviations is a minimum when taken from the arithmetic mean. See p. 350, and reference to Yule.

[3] A term introduced by Sir Francis Galton in his studies of inheritance. As Yule suggests such lines might more fittingly be called "characteristic lines." Yule, G. U., *op. cit.*, p. 177.

(1) the correlation coefficient, $r$, (2) the standard deviation—$\sigma_x$—of the $X$ series, and (3) the standard deviation—$\sigma_y$—of the $Y$ series.

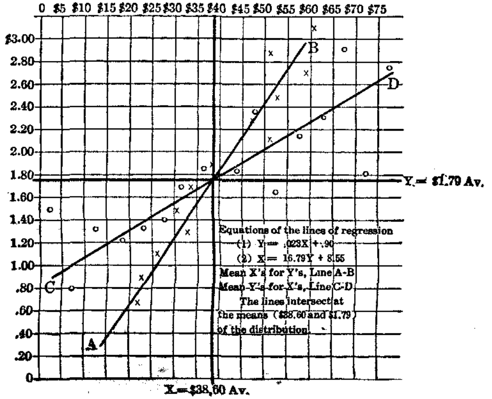
| SERIES $X$ | SERIES $Y$ |
|---|---|
| Rent per 100 sq. ft. of Floor Space | Rent per \$100 of Sales |
| Average = \$38.60 | Average = \$1.79 |
| Standard Deviation = \$18.40 | Standard Deviation = \$ .69 |

$$r = + .63$$

The regression coefficient of $X$—rent per 100 square feet of floor space—on $Y$—rent per \$100 of sales $= r\,\frac{\sigma_x}{\sigma_y}$. Substituting the values above, we get $.63\,\frac{18.4}{.69} = 16.79$. That is, $x = 16.79\ y$.

FIGURE 78

REGRESSION LINES OF RENT PER UNIT OF FLOOR SPACE ON RENT PER UNIT OF SALES, AND RENT PER UNIT OF SALES ON RENT PER UNIT OF FLOOR SPACE FOR 150 RETAIL CLOTHING STORES



What does such a coefficient mean? If stores were selected with rent per $100 of sales, 1 per cent above the average, the regression coefficient, 16.79 of rent per 100 square feet relative to rent per $100 of sales, indicates that we should expect the stores selected to pay about $16.79 above the average amount per 100 square feet of floor space. In general, if stores paying $x$ dollars in rent per $100 of sales above or below the mean were selected, we should expect the amounts which they pay in rent per 100 square feet of floor space to be 16.79 $x$ from the average amount so paid.

The regression coefficient of $Y$—rent per $100 of sales—on $X$—rent per 100 square feet $= r\,\frac{\sigma_y}{\sigma_x} = .63\,\frac{.69}{18.4} = .023$. That is, $y = .023\,x$.

If, for instance, stores were selected which paid in rent per square foot of floor space $10 more than the average, the regression coefficient, .023, indicates that they would most probably pay in rent per $100 of sales .23 per cent above the average.

Lines AB—regression of $X$ on $Y$—and CD—regression of $Y$ on $X$—are drawn in keeping with the respective coefficients of regression, $x = 16.79\,y$; and $y = .023\,x$. The manner in which this is done is by locating two or more points of $X$ on $Y$ and $Y$ on $X$ by the use of the following formulæ:

For $Y$ on $X$

$y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}(x - \bar{x})$, where $y$ and $x =$ any values of the correlated values, and $\bar{y}$ and $\bar{x}$ are the means of the respective series.

Inserting values in this equation we get

$$y - 1.79 = .023\ (x - 38.6)$$

For $X$ on $Y$ the corresponding formula is

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Inserting values in this equation we get

$$x - 38.6 = 16.79\ (y - 1.79)$$

Then solving for values of $y$ with different values of $x$, and for values of $x$ for different values of $y$ we get

| Regressions of $Y$ Series on $X$ Series | | Regressions of $X$ Series on $Y$ Series | |
|---|---|---|---|
| $x$ | $y$ | $y$ | $x$ |
| 30 | 1.59 | 1.00 | 25.34 |
| 40 | 1.82 | 2.00 | 42.13 |
| 50 | 2.05 | 3.00 | 58.92 |
| etc. | etc. | etc. | etc. |

When $x$ increases by 10, $y$ increases by $10 \times .023$ or .23.

When $y$ increases by 1.00, $x$ increases by 16.79.

By using these relations, the AB line—regression of $X$ on $Y$—and the CD line—$Y$ on $X$—are drawn.

The regression coefficient is therefore a fixed ratio between the deviations of attributes in correlated series whereby it is possible, if the amount is known by which the attribute in one series deviates from the mean, to predict the extent to which the associated attribute will most probably deviate from its mean. The extent of deviation in each series is indicated in *its own* unit of measurement. Prediction, of course, rests upon the law of probability and theory of error already discussed.[1]

### *(5) The Probable Error of the Coefficient of Correlation*

Is the amount of negative correlation between the load factor and income per K.W.H., and the amount of positive correlation between rent and sales and rent and floor space "significant"? A similar question was asked [2] about individual measurements and the means of a series of measurements. The answer was found in the probable error concept. It was said that the probable error is a measure which if added to and subtracted from a most probable measurement—mean in the case of an individual measurement; average of a series

[1] For an excellent discussion of regression lines and coefficients see Rugg, H. O., *Statistical Methods Applied to Education*, Houghton Mifflin, 1917, pp. 252-259.

[2] See p. 370 *ff.*

of means for a mean—gives amounts within which the chances are even that an item of the same type, if selected at random, will fall.

The correlation coefficient too has a probable error. It is that amount on either side of the average coefficient of correlation within which half of the values of a large number of coefficients fall if computed from series of pairs of items chosen at random from a universe having in general the given correlation coefficient. That is, if from a large population successive pairs of samples were drawn at random and their correlation coefficients determined, the results would differ. They, however, would tend to describe the normal probability curve, being systematically distributed about a mean. The probable error of $r$, therefore, is an amount which if added to and subtracted from the average correlation coefficient produces amounts within which the chances are even that a coefficient of correlation from a series selected at random will fall.

The formula for the probable error of Pearson's coefficient of correlation—$r$—is $.6745 \frac{1 - r^2}{\sqrt{n}}$, where $n$ is the number of items paired, and $r$ the coefficient itself. The amount secured from this formula is a function of the size of the coefficient—$r$—and the number of items.

It has become conventional to say that for $r$ to be significant it must be at least six times its probable error. Under such circumstances the odds are large that another coefficient computed from series selected at random would fall within a range above and below the mean set by such an amount. Judged by this standard, both correlation coefficients are significant. The coefficient, —0.517, between the load factor and K.W.H. is more than seven times its probable error, .0721. The coefficient for rental payments in terms of sales and in units of floor space, + .63,[1] is approximately twenty times its probable

[1] By another computation it is + .64 — see Table 72.

error, .033. The coefficients with their probable errors written in the customary manner are as follows:

Load factor and K.W.H.: $r = -0.517 \pm .0721$

Rents in units of sales and in units of floor space: $r = +.63 \pm .033$.

## 2. THE CONCURRENT DEVIATION METHOD

If a measure of association in the *direction* of change alone is desired, the method of concurrent deviations may be used.

Table 73 is composed of four primary sections. In the upper left-hand corner the stores which had expenses above the average in the first year[1] and also in the second year are tabulated in classified groups according to per cents by which their expenses exceed the averages in the respective years. The upper right-hand corner contains the stores having expenses greater than the average in the first and less than the average in the second year—the deviations being shown in the same manner as in the quarter just described. Similarly, stores having expenses less than the average in the first and greater than the average in the second year are listed in the lower left-hand corner. The lower right-hand corner contains stores the expenses of which in both years were less than the average. Such an arrangement constitutes a four-part "double frequency" table.

An inspection of the table indicates that stores which had expenses higher or lower than the average in the first year generally had expenses higher or lower than the average in the second year. A few stores, however, the expenses of which were higher or lower than the average in the first year had expenses lower or higher in the second year. In no one of the four sections is there complete identity as to the amount of the difference of the expenses from the average for the stores in the first and in the second year.

[1] By "first" and "second" years are meant the first and second of a pair of years, as 1916 and 1917, 1917 and 1918, etc. Table 73 is the summation of such distributions for the four pairs of years, 1916 to 1920, inclusive.

## TABLE 73

NUMBER OF IDENTICAL RETAIL CLOTHING STORES DISTRIBUTED ACCORDING TO THE AMOUNT AND TYPE OF THEIR EXPENSE DEVIATIONS FROM THE AVERAGE IN TWO SUCCESSIVE YEARS

| Year: First (Number of Stores with Per Cent Deviations from the Average in the First of the Pair of Years) | Second: Position of Item | Groups | Greater than the Average: Total | Greater: 40 & Over | Greater: 30-40 | Greater: 20-30 | Greater: 10-20 | Greater: -10 | Less than the Average: -10 | Less: 10-20 | Less: 20-30 | Less: 30-40 | Less: 40 & Over | Less: Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Greater than the Average | | Total | 266 | 22 | 27 | 54 | 76 | 87 | 50 | 13 | 3 | 4 | | 70 |
| | | 40 & over | 27 | 11 | 4 | 8 | 3 | 1 | | 1 | | | | 1 |
| | | 30-40 | 25 | 2 | 4 | 5 | 8 | 6 | 1 | | | 1 | | 2 |
| | | 20-30 | 60 | 5 | 8 | 18 | 14 | 15 | 3 | 1 | | | | 4 |
| | | 10-20 | 66 | | 4 | 9 | 26 | 27 | 14 | 3 | | 1 | | 18 |
| | | -10 | 88 | 4 | 7 | 14 | 25 | 38 | 32 | 8 | 3 | 2 | | 45 |
| Less than the Average | | -10 | 47 | 1 | 1 | 3 | 10 | 32 | 41 | 26 | 8 | | 1 | 76 |
| | | 10-20 | 17 | | 1 | | 3 | 13 | 20 | 35 | 32 | 6 | 1 | 94 |
| | | 20-30 | 10 | | | 2 | 1 | 7 | 9 | 20 | 26 | 9 | 2 | 66 |
| | | 30-40 | | | | | | | 2 | 2 | 4 | 7 | 6 | 21 |
| | | 40 & over | | | | | | | 1 | | 2 | 4 | 6 | 13 |
| | | Total | 74 | 1 | 2 | 5 | 14 | 52 | 73 | 83 | 72 | 26 | 16 | 270 |

The degree of correlation between the positions of the stores relative to the averages in the first and second years of each pair may be measured by the formula: [1]

[1] If the quantity, $2c-n$, is negative, a minus sign is used before it and before the radical so that the square root can be taken.

$$r = \pm\sqrt{\pm\left(\frac{2c - n}{n}\right)}$$

where $r$ is the coefficient of correlation;
$c$, the number of pairs having like signs; and
$n$, the number of pairs of items.

The association of positions relative to the averages may be summarized as follows:

Second of the Pairs of Years

| First of the Pairs of Years | + | − |
|---|---|---|
| + | 266 | 70 |
| − | 74 | 270 |

Inserting the values into the formula we get

$$r = \pm\sqrt{\pm\frac{2(266 + 270) - 680}{680}} = \sqrt{\frac{+1072 - 680}{680}}$$

$$= \sqrt{\frac{+392}{680}} = \sqrt{+.576} = +.76$$

Pearson's $r$, which measures not only the direction but also the amount of deviation relative to the average, gives a value of $+.74 \pm .012$.

### 3. GRAPHIC METHODS OF SHOWING ASSOCIATION BETWEEN DIFFERENT VARIABLES

Figure 79 shows an inverse relation between the amount of annual sales in retail clothing stores and the size of inventories per unit of sales.

Figure 80 shows a direct relation between the amount of annual sales in retail clothing stores and the annual rates of stock turnover.

FIGURE 79

AMOUNTS OF INVENTORY PER $100 OF TOTAL NET SALES FOR STORES CLASSIFIED BY SIZE, 1919, 1918, AND 1914, COMBINED

| Net Sales (In 000's) | Number of Stores | Inventories Per $100 of Sales: Amounts |
|---|---|---|
| Total (Average) | 920 | $38.00 |
| Under $20 | 50 | 70.67 |
| $20 to $40 | 239 | 53.29 |
| $40 to $60 | 209 | 46.73 |
| $60 to $80 | 126 | 44.53 |
| $80 to $100 | 80 | 41.40 |
| $100 to $140 | 95 | 39.43 |
| $140 to $180 | 43 | 36.67 |
| $180 to $220 | 21 | 29.00 |
| $220 to $300 | 23 | 29.49 |
| $300 to $500 | 22 | 26.57 |
| $500 & over | 12 | 25.75 |
| Under $40 | 289 | 54.97 |
| $40 to $80 | 335 | 45.74 |
| $80 to $180 | 218 | 39.24 |
| $180 & over | 78 | 27.24 |

Figure 81 shows an essentially constant relation between the amount of annual sales in retail clothing stores and the amount paid in wages and salaries as a per cent of total operating expense.

## V. CONCLUSION

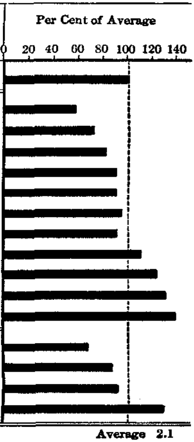
*The discussion of correlation in this chapter* has had to do with its meaning and application under the assumption of the normal law of error distribution. It was in keeping with

such assumption that the Pearsonian coefficient was conceived, and it is only in this connection that the formula accurately measures correlation.

FIGURE 80

ANNUAL RATES OF STOCK TURNOVER FOR STORES CLASSIFIED BY SIZE, 1919

| Net Sales (in 000's) | Number of Stores | Annual Stock Turnover Rates: Rates |
|---|---|---|
| Total (Average) | 314 | 2.1 |
| Under $20 | 3 | 1.2 |
| $20 to $40 | 43 | 1.5 |
| $40 to $60 | 77 | 1.7 |
| $60 to $80 | 45 | 1.9 |
| $80 to $100 | 34 | 1.9 |
| $100 to $140 | 45 | 2.0 |
| $140 to $180 | 22 | 1.9 |
| $180 to $220 | 12 | 2.3 |
| $220 to $300 | 12 | 2.6 |
| $300 to $500 | 14 | 2.7 |
| $500 & over | 7 | 2.9 |
| Under $40 | 46 | 1.4 |
| $40 to $80 | 122 | 1.8 |
| $80 to $180 | 101 | 1.9 |
| $180 & over | 45 | 2.7 |




Bowley's summary of his discussion of correlation may be used to close our own.

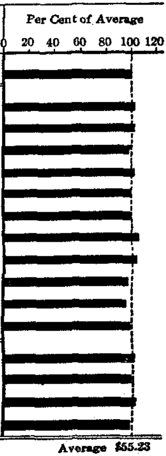
"We may now sum up the treatment of correlation so far. If $(x, y)$ is a pair of measurements (from their averages) of two variables (related in space, in time, in a thing or in an organism), and if when $x$ is given as positive (or negative) there is a presumption that

$y$ is positive (or negative), or a presumption that $y$ is negative (or positive), then the variables are said to be correlated. In such a case $\frac{1}{n}Sxy$ *does not tend to zero when* $n$ *is increased, but to a limit* written as $r\,\sigma_x\,\sigma_y$. $r = 0, = 1, = -1$ have definite meanings; $r$ is sensitive to all kinds of relationship between $x$ and $y$. In general it may be expected to be the greater as $\sigma_a$ (the mean scattering within the arrays) is less. *If* $x$ *and* $y$ *are each the sum of* $p + q$ *independent* elements of which $p$ (only) are common to $x$ and $y$, then $r$ equals $p/(p+q)$, if the standard deviations of the elements are equal. If $x$ and $y$ are generated linearly from a multiplicity of independent

FIGURE 81

Amounts of Wages and Salaries per $100 of Total Expense for Stores Classified by Size, 1919, 1918, and 1914, Combined

| Net Sales (in 000's) | Number of Stores | Wages and Salaries Per $100 of Total Expense: Amounts |
|---|---|---|
| Total (Average) | 929 | $55.23 |
| Under $20 | 48 | 56.30 |
| $20 to $40 | 244 | 55.87 |
| $40 to $60 | 214 | 54.54 |
| $60 to $80 | 130 | 55.85 |
| $80 to $100 | 82 | 55.22 |
| $100 to $140 | 90 | 54.96 |
| $140 to $180 | 44 | 58.26 |
| $180 to $220 | 23 | 57.22 |
| $220 to $300 | 23 | 53.75 |
| $300 to $500 | 21 | 53.20 |
| $500 & over | 10 | 54.87 |
| Under $40 | 292 | 55.92 |
| $40 to $80 | 344 | 55.17 |
| $80 to $180 | 216 | 55.97 |
| $180 & over | 77 | 54.50 |




Average $55.23

causes (some of them common to $x$ and $y$), then $r$ defines the whole frequency distribution of the pairs, the regression loci are rectilinear, and their equations are $y = r\frac{\sigma_y}{\sigma_x}.x$, and $x = r\frac{\sigma_x}{\sigma_y}.y$. If the normal frequency surface cannot be assumed, but regression is rectilinear, the same equation is a good empirical statement of regression. If nothing can be postulated as to the distribution of $x$ and $y$ or the averages of the arrays, the meaning of the numerical value of $r$ is undefined. . . . In general, however, $r$ may be said to measure the amount that is common in the systems of causation of $x$ and $y$."[1]

### References

Bowley, A. L., *Elements of Statistics,* King, London, 1920, Part II, Chapter VI, pp. 350-379.

Bowley, A. L., *Measurement of Groups and Series,* Layton, Loudon, 1903, pp. 61-74, "Correlation between Two Groups"; pp. 82-88, "Correlation between Series."

Davenport, E., *Principles of Breeding,* Ginn and Co., New York, 1907, Chapter 13, pp. 453-472.

Davies, G. R., *Introduction to Economic Statistics,* Century Company, New York, 1922, Chapter VI, pp. 131-148.

Elderton, W. P., *Frequency Curves and Correlation,* Chapter VI, pp. 106-125.

Elderton, W. P. and E. M., *Primer of Statistics,* Black, London, 1910, Chapter 5, pp. 55-72.

Forsyth, C. H., *An Introduction to the Mathematical Analysis of Statistics,* Wiley, New York, 1924, Chapter X, pp. 208-232.

Jones, D. Caradog, *A First Course in Statistics,* Bell, London, 1921, Chapters X, XI, pp. 102-131.

Kelley, Truman, *Statistical Method,* Macmillan & Co., New York, 1923, Chapter VIII, pp. 151-196.

Kincer, J. B., "A Correlation of Weather Conditions and Production of Cotton in Texas," *The Monthly Weather Review,* February, 1915, Vol. 43, pp. 61-65 (U. S. Dept. of Agriculture, Weather Bureau).

King, W. I., *Elements of Statistical Method,* Macmillan, New York, 1912, Chapter XVI, pp. 186-197; Chapter XVII, pp. 197-216.

Mills, Frederick C., *Statistical Methods Applied to Economics and Business,* Holt, New York, 1924, Chapter X, pp. 362-410.

[1] Bowley, A. L., *Elements of Statistics,* 4th Edition, King, London, 1920, pp. 366-367.

MOORE, H. L., *Economic Cycles: Their Law and Cause,* Macmillan, New York, 1914, Chapter V.

PEARL, RAYMOND, *Introduction to Medical Biometry and Statistics,* W. B. Saunders Company, Philadelphia, 1923, Chapter XIV, pp. 292-318.

PEARSON, KARL, *The Grammar of Science,* 3rd Edition, Black, London, 1911, Chapters IV and V.

PERSONS, W. M., "Correlation of Economic Statistics," *Publications of the American Statistical Association,* Vol. 12, 1910, pp. 287-322.

PERSONS, W. M., An "Index of General Business Conditions," *The Review of Economic Statistics,* April, 1919, pp. 130-139.

RIETZ, H. L., and CRATHORNE, A. R., "Simple Correlation," *Handbook of Mathematical Statistics,* Houghton Mifflin, Boston, 1924, pp. 121-138.

RUGG, HAROLD O., *Statistical Methods Applied to Education,* Houghton Mifflin, Boston, 1917, Chapter IX, pp. 233-309.

WHIPPLE, GUY MONTROSE, *Manual of Mental and Physical Tests,* Warwick and York, Baltimore, Chapter III, pp. 14-40, particularly.

YULE, G. U., *Introduction to the Theory of Statistics,* Griffin, London, 1911, Chapters IX, X, and XI, pp. 157-224.

# CHAPTER XIV

## THE TREATMENT AND CORRELATION OF TIME SERIES

### I. Introduction

The graphic representation of time or historical series was discussed in Chapter VIII. In that connection, attention was given primarily to (1) the methods of drawing simple and cumulative graphs, (2) scale conversion, (3) difference vs. ratio charts, (4) simple methods of smoothing time series, etc. Further attention to time series was reserved for this chapter because of the intimate relation of the subject to correlation, and to the discussion which must of necessity precede it. Having now described the different methods of summarizing and comparing statistical series in terms of averages and of measures of dispersion and of skewness; and having stated the concepts of probability, and the theory of error and correlation, we are now ready to discuss the statistical treatment and correlation of time series.

### II. The Nature of Changes in Time Series

The most satisfactory way of showing the changes of a series of data over a period of time is to use a graph or line chart. The time intervals—days, months, years, etc.—are plotted along the abscissa axis, the spacings being proportional to the length of time covered. At the different time units, ordinates are erected according to a scale showing absolute amounts or ratio changes. A line connecting the successive ordinates gives a graphic picture of the ups and downs, in-

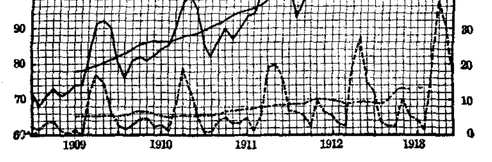
creases, decreases, and general trend which characterize such series. If *nothing more than a general picture of the short- and long-time movements is desired*, smoothed lines drawn free-hand or by a process of averaging will suffice. Indeed, any number of series may be *roughly* compared in this manner. It is only when comparison requires that different types of changes be isolated that more refined methods are needed.

Figure 82 shows the note circulation of chartered Canadian banks and wheat receipts at Fort William and Port Arthur, Canada, from 1909 to 1913, (1) as actual amounts and, (2) as average amounts secured by using a moving average of thirteen months, centered at the seventh month. The lines plotted to the respective averages roughly indicate the trends, while those showing the actual amounts reveal the seasonal changes. Neither of the graphs, however, satisfactorily *measures* the trend or the seasonal movements. More refined methods are necessary.[1]

FIGURE 82

CURVES SHOWING LONG-TIME OR SECULAR CHANGES

(Note Circulation of Canadian Chartered Banks, and Wheat Receipts at Fort William and Port Arthur, Canada, by Months, 1909—1913.)




[1] See the discussion under Section III, *infra*.

The changes in time series may in general terms be spoken of as (1) long-time or secular, and (2) short-time. By a secular change is meant one which characterizes the direction over a number of years. There may be a general tendency for amounts to increase, to decrease, or to assume both directions. The short-time changes are of a periodic or of an irregular type and of relatively short duration.

The *long-time* change may sometimes be generalized into a *trend*, and be represented by a straight line drawn through the data rather than following the movement characterizing the "short" periods. Such a trend line, if positively inclined shows a tendency for the series to increase; if it is negatively inclined, a tendency for it to decrease. The forces back of such long-time trends in series relating to business, industry, social development, etc., are increases in population, improvements in sanitation and health, industrial growth, exhaustion of natural resources, improvements in standards of living, perfection of the arts, and numerous other influences which operate steadily and persistently from year to year.

The *short-time* changes may be classified into three groups: (1) those which are of a seasonal nature, (2) those which are cyclical, and (3) those which may be termed accidental or extraordinary.

The *seasonal* changes are those which are traceable to forces inherent in the seasons themselves. They may be due to meteorological factors such as rainfall and temperature; to demands incident to crop planting, moving and marketing; to fad and fashion in dress; to shifts in population from unfavorable to favorable climates; to conventional practices of debt liquidation, payment of interest on bonds, taking of vacations—in fact to any circumstances peculiar to the seasons as such. Accordingly, in some series they are marked; in others negligible.

By *cyclical* changes are meant those swings in business through periods of expansion, liquidation, depression, and recovery, which have come to be known as "tbe business cycle."

By *accidental* changes or movements are meant those which cannot be traced, (1) to the steady influences of growth or decline, (2) to seasonal adjustments and variations, or (3) to the rhythmical influences of the business cycle. They are rather due to fortuitous events such as wars, strikes, floods, earthquakes, etc.

## III. METHODS OF MEASURING AND ISOLATING TIME CHANGES

Having classified and described the different kinds of changes in time series, the more important methods by which they can be isolated and measured will now be considered.[1]

For the purpose of illustrating different methods, the time series showing the monthly production of pig iron from 1903 to 1916 will be used. The amounts are contained in Table 74.[2]

TABLE 74

MONTHLY PRODUCTION OF PIG IRON IN THE UNITED STATES
(000's of long tons)

| YRS. | JAN. | FEB. | MAR. | APR. | MAY | JUNE | JULY | AUG. | SEPT. | OCT. | NOV. | DEC. | AVE. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | 1472 | 1390 | 1590 | 1608 | 1713 | 1673 | 1546 | 1571 | 1553 | 1425 | 1039 | 846 | 1452 |
| 1904 | 921 | 1205 | 1447 | 1555 | 1534 | 1292 | 1106 | 1167 | 1352 | 1450 | 1486 | 1616 | 1344 |
| 1905 | 1781 | 1597 | 1936 | 1922 | 1963 | 1793 | 1741 | 1843 | 1899 | 2053 | 2014 | 2045 | 1882 |
| 1906 | 2068 | 1904 | 2155 | 2073 | 2098 | 1976 | 2013 | 1926 | 1960 | 2196 | 2187 | 2235 | 2066 |
| 1907 | 2205 | 2045 | 2226 | 2216 | 2295 | 2234 | 2255 | 2250 | 2183 | 2336 | 1828 | 1234 | 2109 |
| 1908 | 1045 | 1077 | 1228 | 1149 | 1165 | 1092 | 1218 | 1348 | 1418 | 1563 | 1577 | 1740 | 1302 |
| 1909 | 1801 | 1703 | 1832 | 1738 | 1880 | 1929 | 2101 | 2246 | 2385 | 2600 | 2547 | 2635 | 2116 |
| 1910 | 2608 | 2397 | 2617 | 2483 | 2390 | 2265 | 2148 | 2106 | 2056 | 2093 | 1909 | 1777 | 2237 |
| *1911* | *1759* | *1794* | *2188* | *2065* | *1893* | *1787* | *1793* | *1926* | *1977* | *2102* | *1999* | *2043* | *1944* |
| 1912 | 2057 | 2100 | 2405 | 2375 | 2512 | 2440 | 2410 | 2512 | 2463 | 2689 | 2630 | 2782 | 2448 |
| 1913 | 2795 | 2586 | 2763 | 2752 | 2822 | 2628 | 2560 | 2543 | 2505 | 2546 | 2233 | 1983 | 2560 |
| *1914* | *1885* | *1888* | *2348* | *2270* | *2093* | *1918* | *1958* | *1995* | *1883* | *1778* | *1518* | *1516* | *1921* |
| 1915 | 1601 | 1675 | 2064 | 2116 | 2203 | 2381 | 2563 | 2780 | 2853 | 3125 | 3037 | 3203 | 2472 |
| 1916 | 3185 | 3087 | 3338 | 3228 | 3351 | 3212 | 3226 | 3204 | 3202 | 3509 | 3312 | 3171 | 3252 |

[1] Much of the following discussion is based upon the work of Professor W. M. Persons, Editor, *The Review of Economic Statistics*, Harvard Economic Service, Cambridge, Mass., to whom all students of the business cycle and of statistical methods are deeply indebted. His unique contributions not only to the methods of isolating the different changes in time series but also to the use of the correlation coefficient in the development of a business barometer and forecaster are outstanding events in the development of statistical methods during the past ten years.

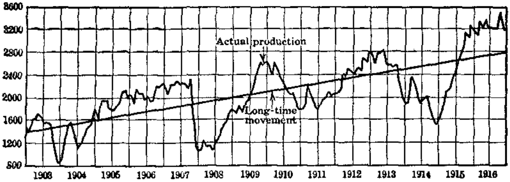
[2] Details taken from *Review of Economic Statistics*, Harvard Committee on Economic Research, January, 1919, p. 66.

Graphic representations of the actual amounts of pig iron produced and of the long-time trend[1] are given in Figure 83.

FIGURE 83

CHART SHOWING THE ACTUAL PRODUCTION OF PIG IRON IN THE UNITED STATES 1903 TO 1916, AND A LINE SHOWING THE LONG-TIME TREND *

(000's of long tons)




* Reproduced by courtesy of the Editors of the *Review of Economic Statistics*, Harvard Committee on Economic Research, Cambridge, Mass.

An inspection of the curve of actual data in Figure 83 shows (1) a long-time tendency for production to increase; (2) more or less periodic rises and falls several years apart; and (3) ups and downs from month to month in each year. The curve seems to contain a definite trend, as well as cyclical and seasonal movements. But what is a high point for one period is a low position for another period, and vice versa. Moreover, the large swings through which the curve passes are blurred by the seasonal changes. It is only by isolating the different movements that a true picture of what happened in production during these years can be secured. Methods of doing this will now be explained.

## 1. *METHODS OF MEASURING LONG-TIME OR SECULAR TREND*

To determine a *trend* in historical data presupposes a *period* for which the trend is to be found. Moreover, the limiting term, "long-time," suggests that the trend is thought of

[1] For the way in which this line is secured see the discussion, pp. 444-447.

as being characteristic—typical or normal—of a period long enough for the influences determining it to work themselves out. Accordingly, the choice of a period requires (1) that as many years as possible should be considered,[1] (2) that periods of evident change in trend be excluded,[2] and (3) that periods of violent change from wars, major strikes, etc.—the "accidental" phases of business growth and decline—be omitted.

The period for which the trend is sought, therefore, cannot be studied too carefully. The addition or the elimination of a year or a number of years may materially change the trend if these conditions are not observed.[3]

From an inspection of Figure 83, it appears that for the production of pig iron in the United States the period 1903-1916 may be used in order to secure a measure of long-time trend.

### (1) *The Free-Hand Method*

A line drawn free-hand through the amounts showing monthly production might serve to give a general notion of the direction of change. *Where* it is to be drawn, however, is a matter of judgment. Different people would draw it at different positions and with varying slopes. If the trend when drawn is to be used as a base from which both seasonal and cyclical variations are to be determined, then its position should not be made a matter of opinion, but so far as

[1] If the trend is to be used in connection with a study of the business cycle, the period should begin and end in the same phase—prosperity, liquidation, depression, recovery—of the cycle.

[2] That is, if a straight line is to be fitted to the data. In some cases some form of a curved line is necessary. Persons' judgment, after having examined a great number of statistical series relating to business and economic phenomena, however, is of interest. "It may be said that for over 95 per cent of economic series it is not worth while to search for a more complicated functional expression between the variables than one of the first degree." (a straight line). Persons, W. M., *Review of Economic Statistics*, April, 1919, p. 135.

[3] See the discussion of this phase of the problem by Persons, W. M., *Review of Economic Statistics*, Harvard Committee on Economic Research, January, 1919, pp. 8-18.

can be of mathematical certainty. Accordingly, other than free-hand methods are necessary, although a line drawn in this fashion may be taken as a first approximation to a line upon which all could agree, and one which would rest upon an acceptable mathematical formula.

### *(2) The Method of Averaging*

A trend line may also be determined by using some form of averaging. But different averages give different results as do also the same averages of different periods. As Persons says, after an exhaustive analysis of the use of moving averages, "It is clear . . . that the use of moving averages does not eliminate the secular trend of the original series. The resulting averages present the problem with which we started, the measurement and elimination of the trend for the period in question."[1]

There is, however, something to be said for the use of moving medians, more particularly when it is certain that the trend does not follow some mathematical law. The medians serve as a first approximation to the line sought, correction from which can be made by some appropriate smoothing device.[2]

### *(3) The Least-Square Method*

The line of "best fit" of a series of points was found in Chapter XIII to be the line from which the sum of the squares of the items, measured parallel to the $Y$ axis, is a minimum.[3] Such a line passes through the arithmetic means

[1] *Op. cit.*, p. 12.

[2] For a defense of the use of the moving median, see King, W. I., "Principles Underlying the Isolation of Cycles and Trends" in *Journal of the American Statistical Association*, December, 1924, pp. 468-475.

[3] The Pearsonian coefficient of correlation is based upon this principle. That is, the slope of the line of regression of $X$ on $Y$ and $Y$ on $X$—when the deviations of the items in each of the series from its arithmetic mean are expressed in units of standard deviation—gives $r$.

of the $X$ and of the $Y$ series—$\frac{\Sigma x}{n}, \frac{\Sigma y}{n}$. The slope—$m$—of the line of regression (also of least squares) is $r\frac{\sigma_y}{\sigma_x}$, where $r$ is the coefficient of correlation, $\sigma_y$ the standard deviation of the $Y$ series, and $\sigma_x$ the standard deviation of the $X$ series. (In time series the $X$ series represents time). But, $r\frac{\sigma_y}{\sigma_x}$ for $m$ reduces to $\frac{\Sigma xy}{\Sigma x^2}$.[1] Therefore, the line of best fit—least squares—may be thought of as the regression line of $Y$ (the items) on $X$ (the time).

Table 75 contains the average monthly totals, the $Y$ series, and the years, the $X$ series, from which the slope of the line in Figure 83 is derived.

The middle point—time—in $X$ is halfway between December, 1909, and January, 1910. The middle amount corresponding to this time is $\frac{29{,}105}{14} = 2078.9$. The annual increment is $\frac{43{,}359}{\frac{910}{2}} = 95.3$. The monthly increment is, therefore, $\frac{95.3}{12} = 7.9$. The annual increment is the amount by which the trend line—see Figure 83—rises from year to year, and the monthly increment, the amount by which it rises from month to month.

Now from the slope—$m = 7.9$ monthly increment—it is only necessary to find the ordinates of the trend. This is done as follows: The middle of the period, 1903-1916, is halfway between December, 1909, and January, 1910. The middle amount corresponding to this period is 2078.9. Accordingly, to get the ordinate for December, 1909, it is necessary to sub-

[1] $r = \frac{\Sigma xy}{n\sigma_1\sigma_2}$. Accordingly, $\frac{\Sigma xy}{n\sigma_x\sigma_y} \times \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{n\sigma_x^2}$. But $\sigma_x = \sqrt{\frac{\Sigma x^2}{n}}$. Accordingly, $\frac{\Sigma xy}{n\sigma_x^2} = \frac{\Sigma xy}{\Sigma x^2}$.

TABLE 75

MONTHLY PRODUCTION OF PIG IRON 1903-1916

(000's tons)

(Showing Method of Determining Monthly Increment of Trend)

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| YEARS $X$ | PRODUCTION MONTHLY $Y$ | DEVIATION IN SERIES * $x$ | DEVIATIONS IN SERIES $X$ SQUARED $x^2$ | $xy$ |
| 1903 | 1452 | —13 | 169 | —18876 |
| 1904 | 1344 | —11 | 121 | —14784 |
| 1905 | 1882 | — 9 | 81 | —16938 |
| 1906 | 2066 | — 7 | 49 | —14482 |
| 1907 | 2109 | — 5 | 25 | —10545 |
| 1908 | 1302 | — 3 | 9 | — 3906 |
| 1909 | 2116 | — 1 | 1 | — 2116 |
| 1910 | 2237 | 1 | 1 | 2237 |
| 1911 | 1944 | 3 | 9 | 5832 |
| 1912 | 2448 | 5 | 25 | 12240 |
| 1913 | 2560 | 7 | 49 | 17920 |
| 1914 | 1921 | 9 | 81 | 17239 |
| 1915 | 2472 | 11 | 121 | 27192 |
| 1916 | 3252 | 13 | 169 | 42276 |
| Total | 29105 | 0 | $\Sigma x^2 = 910$ | $\Sigma xy = 43359$ |

* In order to avoid fractions, since the deviations are taken from the middle of 1909-1910, whole numbers are used, and the $\Sigma x^2$—910—later divided by 2.

tract one half of the monthly increment—that is, $\frac{7.9}{2} = 4$— from 2078.9, which gives 2074.9, or 2075 in round numbers. Then with December, 1909, as a starting point subtract successively the annual increments to get the December ordinates of trend for the previous years, and add them successively to get the December ordinates of trend for the following years.[1]

[1] Of course the trend line can be plotted from any two ordinates as thus determined and the other amounts read directly from the ordinate scale.

It is in this manner that the line of trend in Figure 83 is determined.[1] The actual amounts are shown in column 4 of Table 77.

The line of trend according to the least-square assumption is a "best fit" only for the period to which it applies. The addition of other years or the elimination of some already taken may radically change its position. Moreover, this line can rarely be extended to cover future years because nothing is known about the condition these years will bring. There is no method of stating, as there is in frequency series, the probable dispersion of additional data. As Persons well says:

> "The method of curve-fitting is superior to the method of moving averages for measuring secular trend. The determination of a curve or line which pictures the secular trend of a past period, does not determine present or future trend. The presumption that past trend will continue is strong in some cases and weak in others. The estimate of future trend should be influenced by recent tendencies and current items to some degree, yet we should not lightly conclude from short-time fluctuations that secular trend has changed. . . . The extension of a past trend is a prophecy. It is impossible to get away from that fact. The important thing is that the exact nature of the prophecy be made unmistakable."[2]

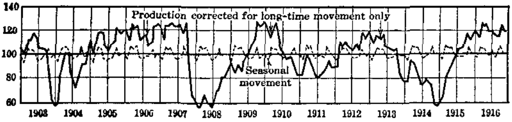
The trend is *eliminated* from the actual items month by month by expressing the items as percentages of the trend. That is, the trend is taken as a base from which the actual items appear as plus (+) or minus (—) deviations. The percentage relations of the items to trend are shown in column 5 of Table 77, and illustrated by the heavy line in Figure 84. This line shows the production (as percentages) corrected for long-time trend.

[1] For a description of another method of determining the annual increment of trend for a straight line which gives the same result as the method of "least squares," see Frickey, Edwin, "The Line of Secular Trend," *The Review of Economic Statistics*, April, 1919, pp. 210-211.

[2] Persons, W. M., *The Review of Economic Statistics*, January, 1919, p. 18.

FIGURE 84

PIG IRON PRODUCTION 1903-1916—FIGURES CORRECTED FOR LONG-TIME TREND (PERCENTAGES) *




* Reproduced by courtesy of the Editors of the *Review of Economic Statistics*, Harvard Committee on Economic Research, Cambridge, Mass.

## 2. METHODS OF MEASURING NORMAL SEASONAL CHANGE

Before measuring seasonal changes, the fact that they exist must first be determined. It is apparent that it is useless to expect a perfect repetition year after year of seasonal swings. Variation characterizes our industrial and social world as it does such pure chance phenomena as dice throws, for instance. Having noted the fact of seasonal change—which may be done from a graphic representation of data—the problem is to secure some measure of the normal or characteristic changes which tend to be repeated year after year. To do this some form of averaging—that is, of reducing detail and variation to type—must be used. But different methods give different results. Which are most satisfactory and why? [1]

[1] The literature on this subject is extensive, and new methods and discussions and criticisms of old ones are constantly appearing. All that can be done in a textbook is to describe briefly the more important methods, and refer students to more detailed and elaborate treatments of the subject. See, for instance, Persons, W. M., "Indices of Business Conditions," *Review of Economic Statistics*, Cambridge, Mass., Jan., 1919, pp. 18-31; King, W. I., "An Improved Method for Measuring the Seasonal Factor" in *Journal of the American Statistical Association*, September, 1924, pp. 301-313; Falkner, H. D., "The Measurement of Seasonal Variation," *Journal of the American Statistical Association*, June, 1924, pp. 167-179 and the literature there referred to.

### (*1*) *Monthly Means or Averages*

If data by months are available over a series of years, and it is desired to get a measure of the normal seasonal variation in the items, the simplest method would appear to be to take an average of some sort of the amounts of the Januaries, the Februaries, etc., and to express them as percentages of their own average. But such a method makes no allowance for the long-time trend, for cyclical movements, nor for accidental disturbances. Moreover, the use of the arithmetic mean gives prominence to the exceptional items, and this is not desired, since what is sought is a picture of the normal seasonal change. This method has little to commend it except the ease with which it may be carried out.[1]

### (*2*) *The Method of Moving Medians*

King [2] has recently suggested a method of measuring the seasonal factor which seems to have considerable merit. The steps in its use are as follows:

a. Plot the original monthly data of the series to be studied.

b. Draw a free-hand curve through the cycles representing as nearly as can be what the data would be if there were no seasonal changes.

c. Read from the curve drawn in "b" the figures each month representing the tentative estimate of the cycle amounts.

d. Divide each of the monthly amounts in "a" by those secured in "c."

e. Take moving medians (King used one covering nine periods) of the percentages for the Januaries, for the Februaries, etc., and plot them to the middle year of the period.

f. Adjust the percentages for the months in each year so that their sum equals twelve.

[1] See Davies, G. R., *Introduction to Economic Statistics*, Century Co., New York, 1922, pp. 116-120 for a discussion of this method, and for a modification of it which eliminates many of its weaknesses.

[2] King, W. I., "An Improved Method of Measuring the Seasonal Factor" in *Journal of the American Statistical Association*, September, 1924, pp. 301-313.

King holds that this method is superior to others because (1) it is easy to understand, (2) can be computed easily, and (3) *gives a separate seasonal index for each year during the period treated.*[1]

### (3) *The Median-Link-Relative Method*[2]

The median-link-relative method of measuring normal seasonal change makes use of an average—the median—and monthly relative numbers calculated on a shifting base. The steps in its use are as follows:

a. From the original monthly items calculate relative or percentage numbers for each month by dividing the amount for each month by the amount for the preceding month and multiplying the result by 100. For instance, $\frac{\text{January}}{\text{December}} \times 100$ gives the January relative; $\frac{\text{February}}{\text{January}} \times 100$, gives the February relative; $\frac{\text{March}}{\text{February}} \times 100$, gives the March relative; and so on through the entire series.

b. Arrange the relative numbers in the form of a frequency table for each pair of months, as $\frac{\text{January}}{\text{December}}, \frac{\text{February}}{\text{January}}$, etc. There will then be as many frequencies for each pair as there are years in the period covered. In the case of pig iron, since the years 1903 to 1916, inclusive, are used there are fourteen. A frequency table arrangement shows the dispersion of the relatives and helps one to decide whether to take a median of all of the items or an average of those near the median. The relatives for pig iron are shown in tabular form in Table 76.

[1] For a discussion of the steps which involve a certain amount of discretion, see King, *op. cit., passim.*

[2] This method was devised by Professor W. M. Persons and is now extensively used. See his discussion of it in comparison with other methods in "Indices of Business Conditions." *Review of Economic Statistics,* Cambridge, Mass., Jan., 1919, pp. 18-31.

c. Inasmuch as a characteristic picture of the dispersion of the relatives is desired, an average least affected by extremes should be used to secure it. Modes would be ideal, but since they are not rigidly defined—indeed, there may be no modes for the series in question—the median of the relatives for each pair of compared months seems most appropriate. The medians for pig iron production are shown at the bottom of Table 76.[1]

TABLE 76

TABLE SHOWING MONTHLY LINK RELATIVES OF PIG IRON PRODUCTION
1903 TO 1916

| YEARS | JAN. / DEC. | FEB. / JAN. | MAR. / FEB. | APR. / MAR. | MAY / APR. | JUNE / MAY | JULY / JUNE | AUG. / JULY | SEPT. / AUG. | OCT. / SEPT. | NOV. / OCT. | DEC. / NOV. | JAN. / DEC. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | 94 | 94 | 114 | 101 | 106 | 98 | 92 | 102 | 99 | 92 | 73 | 81 | |
| 1904 | 109 | 131 | 120 | 107 | 99 | 84 | 86 | 106 | 116 | 107 | 102 | 109 | |
| 1905 | 110 | 90 | 121 | 99 | 102 | 91 | 97 | 106 | 103 | 108 | 98 | 102 | |
| *1906* | *101* | *92* | *115* | *96* | *101* | *94* | *102* | *96* | *102* | *112* | *100* | *102* | |
| 1907 | 99 | 92 | 109 | 100 | 104 | 97 | 101 | 100 | 97 | 107 | 78 | 67 | |
| 1908 | 85 | 103 | 114 | 94 | 101 | 94 | 112 | 112 | 104 | 111 | 101 | 110 | |
| 1909 | 103 | 95 | 107 | 95 | 108 | 103 | 109 | 107 | 106 | 109 | 98 | 103 | |
| 1910 | 99 | 92 | 109 | 95 | 96 | 95 | 95 | 98 | 98 | 102 | 91 | 93 | |
| 1911 | 99 | 101 | 122 | 94 | 92 | 95 | 100 | 107 | 103 | 106 | 95 | 102 | |
| 1912 | 101 | 102 | 114 | 99 | 106 | 97 | 99 | 104 | 98 | 109 | 98 | 106 | |
| *1913* | *100* | *92* | *107* | *100* | *103* | *93* | *97* | *100* | *99* | *102* | *88* | *89* | |
| 1914 | 95 | 100 | 124 | 97 | 92 | 92 | 102 | 102 | 94 | 94 | 85 | 100 | |
| 1915 | 106 | 105 | 123 | 102 | 107 | 105 | 108 | 108 | 103 | 110 | 97 | 105 | |
| 1916 | 100 | 97 | 108 | 97 | 104 | 96 | 100 | 99 | 100 | 110 | 94 | 96 | |
| 1917 | 99 | | | | | | | | | | | | |
| Medians | 100.0 | 96.0 | 114.0 | 98.0 | 102.5 | 95.0 | 100.0 | 103.0 | 101.0 | 107.5 | 96.0 | 102.0 | 100.0 |
| Chain Relatives | 100.0 | 96.0 | 109.4 | 107.2 | 109.9 | 104.4 | 104.4 | 107.5 | 108.6 | 116.8 | 112.1 | 114.3 | 114.3 |
| *Adjusted* | *100.0* | *95.0* | *107.0* | *103.7* | *105.1* | *98.8* | *97.7* | *99.5* | *99.3* | *105.6* | *100.2* | *101.1* | *100.0* |
| Seasonal Indices | 98.9 | 93.9 | 105.9 | 102.6 | 104.0 | 97.7 | 96.6 | 98.4 | 98.3 | 104.5 | 99.2 | 100.0 | |

d. "Chain" the median relatives; that is, successively multiply them together. The amount for $\frac{\text{January}}{\text{December}}$ is taken as 100 and multiplied by the *median* for $\frac{\text{February}}{\text{January}}$, (96.0). This

[1] In some instances, the average of the middle three, or of the middle four items is taken rather than the median item. If no seasonal movement is apparent from the frequency groups, one may sometimes be developed by widening the groups.

gives the *chain relative* of $\frac{\text{February}}{\text{January}}$. This number is then multiplied by the $\frac{\text{March}}{\text{February}}$ *median,* which gives the $\frac{\text{March}}{\text{February}}$ *chain relative.* This process is continued until chain relatives for all of the months are secured. The last multiplication gives the chain relative for December. If there is no secular or long-time trend, the amount for the last $\frac{\text{January}}{\text{December}}$ will be the same as for the first $\frac{\text{January}}{\text{December}}$. If the trend is downward, the last item will be less than the first; if it is upward, it will be more. The method of treating this deficit or excess (as in the case of pig iron)—see "chain relatives" bottom of Table 76—is described in the paragraph immediately following.

e. Since the medians and chain relatives are taken as typical of the entire period, the excess, 14.3 per cent, may be regarded as the average trend. This must be distributed over the 12 monthly relatives. Since the chain relatives were secured by successively multiplying together the medians, any error in seasonal change due to the trend is cumulated from month to month during the year. Accordingly, the excess must be spread over the different months. This may be done arithmetically or geometrically, the latter basis being used to secure the "adjusted relatives" in Table 76.[1]

[1] If the error in the median link relatives is $d$ and the new January chain relative is $A$ (expressed as a decimal—in this case 1.143) then

$$A = (1 + d)^{12}$$

The value of the amount to be distributed may be found from this equation by the use of logarithms. The January chain relative is unaffected by this adjustment. The one for February is divided by $(1 + d)$; the one for March, by $(1 + d)^2$; the one for April by $(1 + d)^3$; and so on, the one for December being divided by $(1 + d)^{11}$. The new January is 100—that is, its excess has been distributed geometrically over the preceding eleven months.

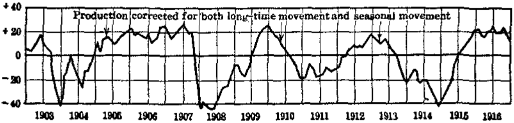
Arithmetically, $\frac{1}{12}$ of the discrepancy should be deducted from the January relative; $\frac{2}{12}$ from the February relative, and so on giving 100 as the relative for the new January on December.

f. The adjusted relatives secured in step "e" are in terms of January as a base. The last step is to express them in terms of the average for the year as a base. This is done by dividing each of them by $\frac{1}{12}$ of the total of the twelve items and multiplying by 100. In this form they are given in the last line of Table 76. These are the *adjusted monthly indexes of seasonal variation.* These indexes are plotted as the broken line above and below the base 100 in Figure 84. They are inserted in column 6 of Table 77.

The seasonal variation may be *eliminated* from a series by subtracting the seasonal indexes month by month each year from the percentage ratios of the actual items to the ordinates of trend. (See Table 77 column 7 which contains the differences taken to the nearest per cent.) A graphic representation of the original data of pig iron production, after they are corrected for both long-time trend and seasonal variation, is shown in Figure 85. The line in this chart—plotted as percentage deviations from a zero or no change line—therefore, represents the cyclical changes (plus the accidental variations) in this series. Technically, it is known as the "Line of Per-

FIGURE 85

Pig Iron Production—Percentages—Corrected for both Secular Trend and Seasonal Variation *




* Reproduced by the courtesy of the Editors of the *Review of Economic Statistics,* Harvard Committee on Economic Research, Cambridge, Mass.

TABLE 77

TABLE SHOWING ACTUAL PIG IRON PRODUCTION, LEAST-SQUARE ORDINATES OF TREND, SEASONAL VARIATION AND CYCLE PERCENTAGES, 1903 TO 1916; AND CYCLE PERCENTAGES OF INTEREST RATES ON 60-90 DAY COMMERCIAL PAPER, NEW YORK, 1903-1916

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| YEAR | MONTH | PIG IRON PRODUCTION | | | | | | CYCLE PER CENTS OF INTEREST RATES ON 60-90 COMMERCIAL PAPER, NEW YORK, 1903-1916 |
| | | Production (000's of tons) | Trend (000's of tons) | Per Cent of Trend 3÷4 ; % | Seasonal Variation % | Cyclical Variations 5—6 ; % | Cycle Per Cents 7 ÷ σ (19.1) | |
| | Jan. | 1472 | 1416 | 104.0 | 98.9 | 5 | .3 | — .1 |
| | Feb. | 1390 | 1424 | 97.6 | 93.9 | 4 | .2 | .1 |
| | Mar. | 1590 | 1432 | 111.0 | 105.9 | 5 | .3 | .5 |
| | Apr. | 1608 | 1440 | 111.7 | 102.6 | 9 | .5 | .2 |
| | May | 1713 | 1448 | 118.3 | 104.0 | 14 | .7 | — .1 |
| 1903 | June | 1673 | 1456 | 114.9 | 97.7 | 17 | .9 | .5 |
| | July | 1546 | 1463 | 105.7 | 96.6 | 9 | .5 | .4 |
| | Aug. | 1571 | 1471 | 106.8 | 98.4 | 8 | .5 | .5 |
| | Sept. | 1553 | 1479 | 105.0 | 98.3 | 7 | .4 | .3 |
| | Oct. | 1425 | 1487 | 95.8 | 104.5 | — 9 | — .5 | — .1 |
| | Nov. | 1039 | 1495 | 69.5 | 99.2 | —30 | —1.6 | .3 |
| | Dec. | 846 | 1503 | 56.3 | 100.0 | —44 | —2.3 | .0 |
| | Jan. | 921 | 1511 | 61.0 | 98.9 | —38 | —2.0 | — .3 |
| | Feb. | 1205 | 1519 | 79.3 | 93.9 | —15 | — .8 | .0 |
| | Mar. | 1447 | 1527 | 94.8 | 105.9 | —11 | — .6 | — .3 |
| | Apr. | 1555 | 1535 | 101.3 | 102.6 | — 1 | — .1 | — .7 |
| | May | 1534 | 1543 | 99.4 | 104.0 | — 5 | — .3 | — .8 |
| 1904 | June | 1292 | 1551 | 83.3 | 97.7 | —14 | — .8 | —1.0 |
| | July | 1106 | 1559 | 70.9 | 96.6 | —26 | —1.4 | —1.3 |
| | Aug. | 1167 | 1567 | 74.5 | 98.4 | —24 | —1.3 | —1.5 |
| | Sept. | 1352 | 1575 | 85.8 | 98.3 | —12 | — .6 | —1.4 |
| | Oct. | 1450 | 1583 | 91.6 | 104.5 | —13 | — .7 | —1.3 |
| | Nov. | 1486 | 1591 | 93.4 | 99.2 | — 6 | — .3 | —1.4 |
| | Dec. | 1616 | 1598 | 101.1 | 100.0 | 1 | .1 | —1.4 |

centage Deviations of Original Items from Secular Trend Corrected for Seasonal Variation."[1]

### 3. CYCLICAL FLUCTUATIONS

The original data, although corrected for trend and seasonal variations as shown in Figure 85, still contain the fluctuations which are due to accidental and fortuitous circumstances. While in a particular case, some satisfactory method might be determined for measuring and removing them too, it is useless to attempt to derive a method which will be generally applicable.[2] Accordingly, cycle percentages, determined in the manner discussed above, represent *true* cycles only when they do not contain accidental fluctuations.

The cyclical fluctuations of time series differ in two major respects: (1) in the amplitude or extent of the variations, and (2) in the time of their occurrence. If the cycles in two series are to be compared, therefore, both of these differences must be taken into account. The ways in which this is done are of interest.

The percentage deviations of cyclical fluctuations in two or more time series may be reduced to a comparable basis by dividing them item by item by the standard deviation of the series to which they belong. This measure of dispersion reduces them to a common denominator in the same way that it does the deviations of items from their respective averages.[3] Such percentages, called "cycles," may then be plotted on a common scale in units of standard deviations. When this is done, the extent or degree of fluctuation through-

[1] An expression found in the writings of Professor Persons, who worked out the above method, and employed in the various studies of the Harvard Committee on Economic Research, Cambridge, Mass.

[2] See Persons, W. M., "An Index of General Business Conditions," *The Review of Economic Statistics*, April, 1919, pp. 137-138, wherein a method of isolating the irregular fluctuations for the value of building permits, 1903-1916, is worked out.

[3] See the discussion of the coefficient of dispersion based on the standard deviation, *supra*, p. 355.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| Year | Month | Pig Iron Production | | | | | | Cycle Per Cents of Interest Rates on 60-90 Commercial Paper, New York, 1903-1916 |
| | | Production (000's of tons) | Trend (000's of tons) | Per Cent of Trend 3 ÷ 4 ; % | Seasonal Variation % | Cyclical Variations 5—6 ; % | Cycle Per Cents 7 ÷ σ (19.1) | |
| 1905 | Jan. | 1781 | 1606 | 110.9 | 98.9 | 12 | .6 | —1.1 |
| | Feb. | 1597 | 1614 | 98.9 | 93.9 | 5 | .3 | — .9 |
| | Mar. | 1936 | 1622 | 119.4 | 105.9 | 13 | .7 | —1.0 |
| | Apr. | 1922 | 1630 | 117.9 | 102.6 | 15 | .8 | — .8 |
| | May | 1963 | 1638 | 118.2 | 104.0 | 16 | .8 | — .7 |
| | June | 1793 | 1646 | 108.9 | 97.7 | 11 | .6 | — .8 |
| | July | 1741 | 1654 | 105.3 | 96.6 | 9 | .5 | — .7 |
| | Aug. | 1843 | 1662 | 110.9 | 98.4 | 13 | .7 | —1.0 |
| | Sept. | 1899 | 1670 | 113.7 | 98.3 | 15 | .8 | — .9 |
| | Oct. | 2053 | 1678 | 122.3 | 104.5 | 18 | .9 | — .8 |
| | Nov. | 2014 | 1686 | 121.1 | 99.2 | 20 | 1.0 | .1 |
| | Dec. | 2045 | 1694 | 120.7 | 100.0 | 21 | 1.1 | .2 |
| 1906 | Jan. | 2068 | 1702 | 121.5 | 98.9 | 23 | 1.2 | .1 |
| | Feb. | 1904 | 1710 | 111.3 | 93.9 | 17 | .9 | .4 |
| | Mar. | 2155 | 1718 | 125.4 | 105.9 | 20 | 1.0 | .5 |
| | Apr. | 2073 | 1726 | 120.1 | 102.6 | 18 | .9 | .8 |
| | May | 2098 | 1733 | 121.1 | 104.0 | 17 | .9 | .8 |
| | June | 1976 | 1741 | 113.5 | 97.7 | 16 | .8 | .8 |
| | July | 2013 | 1749 | 115.1 | 96.6 | 18 | .9 | .8 |
| | Aug. | 1926 | 1757 | 109.6 | 98.4 | 11 | .6 | .9 |
| | Sept. | 1960 | 1765 | 111.0 | 98.3 | 13 | .7 | 1.1 |
| | Oct. | 2196 | 1773 | 123.9 | 104.5 | 19 | 1.0 | .8 |
| | Nov. | 2187 | 1781 | 122.8 | 99.2 | 24 | 1.3 | .9 |
| | Dec. | 2235 | 1789 | 124.9 | 100.0 | 25 | 1.3 | .8 |
| 1907 | Jan. | 2205 | 1797 | 122.7 | 98.9 | 24 | 1.3 | 1.3 |
| | Feb. | 2045 | 1805 | 113.3 | 93.9 | 19 | 1.0 | 1.4 |
| | Mar. | 2226 | 1813 | 122.8 | 105.9 | 17 | .9 | 1.5 |
| | Apr. | 2216 | 1821 | 121.7 | 102.6 | 19 | 1.0 | 1.3 |
| | May | 2295 | 1829 | 125.5 | 104.0 | 21 | 1.1 | .9 |
| | June | 2234 | 1837 | 121.6 | 97.7 | 24 | 1.3 | 1.2 |
| | July | 2255 | 1845 | 122.2 | 96.6 | 26 | 1.4 | 1.1 |
| | Aug. | 2250 | 1853 | 121.4 | 98.4 | 23 | 1.2 | 1.3 |
| | Sept. | 2183 | 1861 | 117.3 | 98.3 | 19 | 1.0 | 1.5 |
| | Oct. | 2336 | 1868 | 125.1 | 104.5 | 20 | 1.1 | 1.7 |
| | Nov. | 1828 | 1876 | 97.4 | 99.2 | — 2 | — .1 | 2.2 |
| | Dec. | 1234 | 1884 | 65.5 | 100.0 | —34 | —1.8 | 2.7 |

out the *different* phases of the cycle become directly comparable.

The cycle percentages for pig iron production, 1903 to 1916, are found in column 8 of Table 77. In this case each of the percentages in column 7 has been divided by 19.1—the standard deviation for this series.

But the "timing" of cyclical fluctuations in different series varies. If it is desired to compare both the amplitude and congruence of change then the method of correlation must be used. A discussion of this phase of the subject immediately follows.

## IV. The Correlation of Time Series

The distinction between correlation and narrow causation was fully developed in Chapter XIII. Nothing further needs to be said about it here except again to call attention to the fact that comparisons generally involve some idea of establishing causation or correlation. Now, the characteristic thing about time series is that the items are "ordered in time," to use Professor Persons' phrase. The relations of the items one to the other as thus ordered are due, among other things, to long-time and short-time influences of a variety of types. Accordingly, if the degree of association between historical series is the object sought by comparison, it is useless to correlate them until, so far as is possible, the different types of fluctuations have been isolated. "It is of little avail (or actually misleading) to compute the coefficient of correlation from pairs of actual items. In case the two series possess definite trends, or seasonal variation the coefficient of correlation for the items will yield a value different from zero. Having found such a coefficient we would be unable to say what contributed most largely to the result—similar (or diverse) trends, seasonal variations, cyclical movements, or irregular fluctuations." [1]

[1] Persons, W. M., "Correlation of Time Series" in *Handbook of Mathematical Statistics*, H. L. Rietz, Editor in Chief, Houghton Mifflin, Boston, 1924, pp. 150-151.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| Year | Month | Pig Iron Production | | | | | | Cycle Per Cents of Interest Rates on 60-90 Commercial Paper, New York, 1903-1916 |
| | | Production (000's of tons) | Trend (000's of tons) | Per Cent of Trend 3 ÷ 4 ; % | Seasonal Variation % | Cyclical Variations 5—6 ; % | Cycle Per Cents 7 ÷ σ (19.1) | |
| 1908 | Jan. | 1045 | 1892 | 55.2 | 98.9 | —44 | —2.3 | 1.9 |
| | Feb. | 1077 | 1900 | 56.7 | 93.9 | —37 | —2.0 | .6 |
| | Mar. | 1228 | 1908 | 64.4 | 105.9 | —42 | —2.2 | 1.0 |
| | Apr. | 1149 | 1916 | 60.0 | 102.6 | —43 | —2.2 | — .2 |
| | May | 1165 | 1924 | 60.6 | 104.0 | —44 | —2.3 | — .5 |
| | June | 1092 | 1932 | 56.5 | 97.7 | —41 | —2.1 | — .7 |
| | July | 1218 | 1940 | 62.8 | 96.6 | —34 | —1.8 | — .9 |
| | Aug. | 1348 | 1948 | 69.2 | 98.4 | —29 | —1.5 | —1.4 |
| | Sept. | 1418 | 1956 | 72.5 | 98.3 | —26 | —1.4 | —1.5 |
| | Oct. | 1563 | 1964 | 79.6 | 104.5 | —25 | —1.3 | —1.3 |
| | Nov. | 1577 | 1972 | 80.0 | 99.2 | —19 | —1.0 | —1.2 |
| | Dec. | 1740 | 1980 | 87.9 | 100.0 | —12 | — .6 | —1.6 |
| 1909 | Jan. | 1801 | 1988 | 90.6 | 98.9 | — 8 | — .4 | —1.1 |
| | Feb. | 1703 | 1996 | 85.3 | 93.9 | — 8 | — .4 | — .9 |
| | Mar. | 1832 | 2003 | 91.5 | 105.9 | —14 | — .8 | —1.1 |
| | Apr. | 1738 | 2011 | 86.4 | 102.6 | —16 | — .8 | —1.0 |
| | May | 1880 | 2019 | 93.1 | 104.0 | —11 | — .6 | —1.0 |
| | June | 1929 | 2027 | 95.2 | 97.7 | — 3 | — .2 | —1.0 |
| | July | 2101 | 2035 | 103.2 | 96.6 | 7 | .4 | —1.2 |
| | Aug. | 2246 | 2043 | 109.9 | 98.4 | 12 | .6 | — .9 |
| | Sept. | 2385 | 2051 | 116.3 | 98.3 | 18 | .9 | —1.0 |
| | Oct. | 2600 | 2059 | 126.3 | 104.5 | 22 | 1.1 | — .3 |
| | Nov. | 2547 | 2067 | 123.2 | 99.2 | 24 | 1.3 | .0 |
| | Dec. | 2635 | 2075 | 127.0 | 100.0 | 27 | 1.4 | — .2 |
| 1910 | Jan. | 2608 | 2083 | 125.2 | 98.9 | 26 | 1.4 | .2 |
| | Feb. | 2397 | 2091 | 114.6 | 93.9 | 20 | 1.1 | .2 |
| | Mar. | 2617 | 2099 | 124.7 | 105.9 | 19 | 1.0 | .0 |
| | Apr. | 2483 | 2107 | 117.8 | 102.6 | 15 | .8 | .4 |
| | May | 2390 | 2115 | 113.0 | 104.0 | 9 | .5 | .5 |
| | June | 2265 | 2123 | 106.7 | 97.7 | 9 | .5 | .8 |
| | July | 2148 | 2131 | 100.8 | 96.6 | 4 | .2 | 1.1 |
| | Aug. | 2106 | 2138 | 98.5 | 98.4 | 0 | .0 | .7 |
| | Sept. | 2056 | 2146 | 95.8 | 98.3 | — 3 | — .2 | .5 |
| | Oct. | 2093 | 2154 | 97.2 | 104.5 | — 7 | — .4 | .5 |
| | Nov. | 1909 | 2162 | 88.3 | 99.2 | —11 | — .6 | .6 |
| | Dec. | 1777 | 2170 | 81.9 | 100.0 | —18 | — .9 | — .5 |

Bowley has stated the same thought as follows:

"If we take two things which are absolutely disconnected, except that they are both phenomena arising in the progress of society, and work out the coefficient by the straightforward rule, we shall find there is some correlation. If two curves have short fluctuations which are correlated, but opposite symptoms, then owing to the symptom apart from the fluctuations there would be negative correlation, while owing to the fluctuations apart from the symptom there would be positive correlation; and when both are taken into account the correlation may be positive, zero, or negative."[1]

If this is true, then correlation or association is best measured by using series from which both the trend and the seasonal variation have been eliminated. The cycle percentages are distinctly less "ordered in time" than are the original items. Series relating to business and economic phenomena are much more alike in their cyclical relations alone than they are in all of their fluctuations. Their trends and seasonal variations are peculiar to themselves; their cyclical fluctuations are the results of underlying business conditions affecting industry and trade generally.

Two methods are available for correlating the cyclical variations of two or more series; (1) the graphic method, and (2) the use of the Pearsonian coefficient. The graphic method indicates the fact of correlation, but it does not measure it. Pearson's $r$ does both. Moreover, the graphic method of superimposing one "corrected" series over the other roughly indicates the appropriate period of lag which will give the highest degree of correlation. It does not, however, measure the correlation for different "timings." This is done only by the use of the numerical measure of correlation—Pearson's $r$. How is this measure applied to "cycle percentages"?

The different steps in correcting original items for secular trend and seasonal variation, as outlined above for pig iron production, give a series of percentages. In order to make them comparable, it has been found to be appropriate to divide

[1] Bowley, A. L., *Measurement of Groups and Series*, Layton, London, 1903, p. 83.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| Year | Month | Pig Iron Production | | | | | | Cycle Per Cents of Interest Rates on 60-90 Commercial Paper, New York, 1903-1916 |
| | | Production (000's of tons) | Trend (000's of tons) | Per Cent of Trend 3 ÷ 4 ; % | Seasonal Variation % | Cyclical Variations 5—6 ; % | Cycle Per Cents 7 ÷ σ (19.1) | |
| 1911 | Jan. | 1759 | 2178 | 80.8 | 98.9 | —18 | — .9 | — .6 |
| | Feb. | 1794 | 2186 | 82.1 | 93.9 | —12 | — .6 | — .2 |
| | Mar. | 2188 | 2194 | 99.7 | 105.9 | — 6 | — .3 | — .6 |
| | Apr. | 2065 | 2202 | 93.8 | 102.6 | — 9 | — .5 | — .7 |
| | May | 1893 | 2210 | 85.7 | 104.0 | —18 | — .9 | — .6 |
| | June | 1787 | 2218 | 80.6 | 97.7 | —17 | — .9 | — .4 |
| | July | 1793 | 2226 | 80.5 | 96.6 | —16 | — .8 | — .6 |
| | Aug. | 1926 | 2234 | 86.2 | 98.4 | —12 | — .6 | — .6 |
| | Sept. | 1977 | 2242 | 88.2 | 98.3 | —10 | — .5 | — .5 |
| | Oct. | 2102 | 2250 | 93.4 | 104.5 | —11 | — .6 | — .8 |
| | Nov. | 1999 | 2258 | 88.5 | 99.2 | —11 | — .6 | —1.1 |
| | Dec. | 2043 | 2266 | 90.2 | 100.0 | —10 | — .5 | — .5 |
| 1912 | Jan. | 2057 | 2273 | 90.5 | 98.9 | — 8 | — .5 | — .6 |
| | Feb. | 2100 | 2281 | 92.1 | 93.9 | — 2 | — .1 | — .4 |
| | Mar. | 2405 | 2289 | 105.1 | 105.9 | — 1 | — .1 | — .1 |
| | Apr. | 2375 | 2297 | 103.4 | 102.6 | 1 | .1 | — .1 |
| | May | 2512 | 2305 | 109.0 | 104.0 | 5 | .3 | .1 |
| | June | 2440 | 2313 | 105.5 | 97.7 | 8 | .4 | .1 |
| | July | 2410 | 2321 | 103.8 | 96.6 | 7 | .4 | .4 |
| | Aug. | 2512 | 2329 | 107.9 | 98.4 | 9 | .5 | .5 |
| | Sept. | 2463 | 2337 | 105.4 | 98.3 | 7 | .4 | .8 |
| | Oct. | 2689 | 2345 | 114.7 | 104.5 | 10 | .5 | 1.1 |
| | Nov. | 2630 | 2353 | 111.8 | 99.2 | 13 | .7 | 1.1 |
| | Dec. | 2782 | 2361 | 117.8 | 100.0 | 18 | .9 | 1.3 |
| 1913 | Jan. | 2795 | 2369 | 118.0 | 98.9 | 19 | 1.0 | .7 |
| | Feb. | 2586 | 2377 | 108.8 | 93.9 | 15 | .8 | 1.0 |
| | Mar. | 2763 | 2385 | 115.8 | 105.9 | 10 | .5 | 1.8 |
| | Apr. | 2752 | 2393 | 115.0 | 102.6 | 12 | .7 | 1.6 |
| | May | 2822 | 2401 | 117.5 | 104.0 | 14 | .7 | 1.5 |
| | June | 2628 | 2408 | 109.1 | 97.7 | 11 | .6 | 2.3 |
| | July | 2560 | 2416 | 106.0 | 96.6 | 9 | .5 | 2.2 |
| | Aug. | 2543 | 2424 | 104.9 | 98.4 | 6 | .4 | 1.7 |
| | Sept. | 2505 | 2432 | 103.0 | 98.3 | 5 | .3 | 1.2 |
| | Oct. | 2546 | 2440 | 104.4 | 104.5 | 0 | .0 | 1.0 |
| | Nov. | 2233 | 2448 | 91.2 | 99.2 | — 8 | — .4 | 1.0 |
| | Dec. | 1983 | 2456 | 80.7 | 100.0 | —19 | —1.0 | 1.0 |

them by the standard deviation of the series to which they belong. In this form, they are multiples of this common divisor. Accordingly, to correlate them with another series similarly corrected it is necessary only to multiply together the corresponding deviations in the two series, algebraically sum or total the products and divide by the number of paired items involved. This follows because (1) in each of two series *the algebraic sum of the deviations from the line of secular* trend equals or closely approximates zero,[1] and (2) the cycle percentages are themselves expressed in units of standard deviations. Accordingly, the formula, $r = \frac{\Sigma\, x\, y}{n\, \sigma_1\, \sigma_2}$, for original data, becomes $\frac{\Sigma\, xy}{n}$ for cycle percentages.

The cycle percentages for interest rates on 60-90 day commercial paper in New York [2] are shown in Table 77 column 9. If these two series are correlated by pairing corresponding months—that is, by multiplying the (.3) for January, 1903, pig iron production in column 8 of Table 77 by the (—.1) for January, 1903, 60-90 day interest rate, in column 9; the February (.2) by the February (.1); and so on for the remainder of the months during 1903 to 1916—the eorrelation coefficient $r$ is found to be + .109. If coefficients are worked out with interest rates lagged after pig iron production, different results will be secured. If interest rates are lagged 4 months —*that is, if May, 1903, interest rate cycles are paired with* January, 1903, pig iron production cycles, June with February and so on—the correlation is + .50. Successive lagging of interest rates gives the following coefficients: 5 months, + .52; 6 months, + .57; 7 months, + .58; 8 months, + .57; 9 months, + .57; 10 months, + .55. Accordingly, maximum correlation

[1] The actual deviations will always equal zero, and the percentages closely approximate it in most cases.

[2] Data are taken from the *Review of Economic Statistics*, January, 1919, p. 122. They are secured in the same manner as the corresponding data for pig iron production.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| Year | Month | Pig Iron Production | | | | | | Cycle Per Cents of Interest Rates on 60-90 Commercial Paper, New York, 1903-1916 |
| | | Production (000's of tons) | Trend (000's of tons) | Per Cent of Trend 3 ÷ 4 ; % | Seasonal Variation % | Cyclical Variations 5—6 ; % | Cycle Per Cents 7 ÷ σ (19.1) | |
| | Jan. | 1885 | 2464 | 76.5 | 98.9 | —22 | —1.2 | .3 |
| | Feb. | 1888 | 2472 | 76.4 | 93.9 | —18 | — .9 | — .2 |
| | Mar. | 2348 | 2480 | 94.7 | 105.9 | —11 | — .6 | — .3 |
| | Apr. | 2270 | 2488 | 91.2 | 102.6 | —11 | — .6 | — .4 |
| | May | 2093 | 2496 | 83.9 | 104.0 | —20 | —1.0 | — .1 |
| 1914 | June | 1918 | 2504 | 76.6 | 97.7 | —21 | —1.1 | .1 |
| | July | 1958 | 2512 | 78.0 | 96.6 | —19 | —1.0 | .4 |
| | Aug. | 1995 | 2520 | 79.2 | 98.4 | —19 | —1.0 | 2.3 |
| | Sept. | 1883 | 2528 | 74.5 | 98.3 | —24 | —1.3 | 2.4 |
| | Oct. | 1778 | 2536 | 70.1 | 104.5 | —34 | —1.8 | 2.0 |
| | Nov. | 1518 | 2543 | 59.7 | 99.2 | —40 | —2.1 | 1.1 |
| | Dec. | 1516 | 2551 | 59.4 | 100.0 | —41 | —2.1 | — .5 |
| | Jan. | 1601 | 2559 | 62.6 | 98.9 | —36 | —1.9 | — .4 |
| | Feb. | 1675 | 2567 | 65.3 | 93.9 | —29 | —1.5 | — .2 |
| | Mar. | 2064 | 2575 | 80.2 | 105.9 | —25 | —1.3 | — .8 |
| | Apr. | 2116 | 2583 | 81.9 | 102.6 | —21 | —1.1 | — .4 |
| | May | 2263 | 2591 | 87.3 | 104.0 | —17 | — .9 | — .2 |
| 1915 | June | 2381 | 2599 | 91.6 | 97.7 | — 6 | — .3 | — .1 |
| | July | 2563 | 2607 | 98.3 | 96.6 | 2 | .1 | — .9 |
| | Aug. | 2780 | 2615 | 106.3 | 98.4 | 8 | .4 | —1.0 |
| | Sept. | 2853 | 2623 | 108.8 | 98.3 | 10 | .5 | —1.6 |
| | Oct. | 3125 | 2631 | 118.8 | 104.5 | 14 | .7 | —1.8 |
| | Nov. | 3037 | 2639 | 115.1 | 99.2 | 16 | .8 | —1.8 |
| | Dec. | 3203 | 2647 | 121.0 | 100.0 | 21 | 1.1 | —1.8 |
| | Jan. | 3185 | 2655 | 120.0 | 98.9 | 21 | 1.1 | —1.2 |
| | Feb. | 3087 | 2663 | 115.9 | 93.9 | 22 | 1.1 | — .8 |
| | Mar. | 3338 | 2671 | 125.0 | 105.9 | 19 | 1.0 | —1.0 |
| | Apr. | 3228 | 2678 | 120.5 | 102.6 | 18 | .9 | — .9 |
| | May | 3351 | 2686 | 124.8 | 104.0 | 21 | 1.1 | — .8 |
| 1916 | June | 3212 | 2694 | 119.2 | 97.7 | 21 | 1.1 | — .1 |
| | July | 3226 | 2702 | 119.4 | 96.6 | 23 | 1.2 | .1 |
| | Aug. | 3204 | 2710 | 118.2 | 98.4 | 20 | 1.0 | — .6 |
| | Sept. | 3202 | 2718 | 117.8 | 98.3 | 20 | 1.0 | —1.4 |
| | Oct. | 3509 | 2726 | 128.7 | 104.5 | 24 | 1.3 | —1.5 |
| | Nov. | 3312 | 2734 | 121.1 | 99.2 | 22 | 1.1 | —1.1 |
| | Dec. | 3171 | 2742 | 115.6 | 100.0 | 16 | .8 | — .8 |

occurs when interest rates are lagged seven months after pig iron production. This is the time interval (in monthly units) of "best fit" between cycles of interest rates on 60-90 day commercial paper and cycles of production of pig iron for the period 1903 to 1916.

But different correlation coefficients would be secured if a different period of time—as for instance 1903 to 1914—were used.[1] Indeed, the size of the coefficient is of value for determining not only the best fitting lag but also the best fitting total period for which to correlate the cycle percentages.

Moreover, the coefficients of correlation of cycle percentages of a great number of time series may be used as a basis for selecting those which lag behind or precede other series. It was by their use that Professor Persons originally constructed from the annual data of a large number of statistical series both a business barometer and a forecaster.[2] The same method, elaborated and refined, when applied to data for the pre-war period, 1903 to 1914, laid the foundation for the present business barometric and forecasting lines of the Index of General Business Conditions now currently issued by the Harvard Committee on Economic Research, and described later.[3]

[1] For the coefficients for different periods of lag, see Persons, W. M., "Correlation of Time Series" in Rietz, H. L. (Editor in Chief) *Handbook of Mathematical Statistics*, Houghton Mifflin, 1924, pp. 162-163.

[2] Persons, W. M., "Construction of a Business Barometer Based Upon Annual Data," *American Economic Review*, December, 1916, pp. 739-769.

[3] See *infra*, pp. 538-541. For a complete explanation of the method see Persons, W. M., "Indices of Business Conditions," *Review of Economic Statistics*, January and April, 1919, *passim;* Persons, W. M., "A Non-Technical Explanation of the Index of General Business Conditions," *Review of Economic Statistics*, February, 1920, pp. 39-48; "The Harvard Index of General Business Conditions—Its Interpretation," *Harvard Committee on Economic Research*, 1923 (published separately); "The Revised Index of General Business Conditions," *Review of Economic Statistics*, July, 1923, pp. 187-195.

## V. The Probable Error of the Correlation Coefficient of Time Series [1]

Having computed the correlation coefficient for two series of random samples on the assumptions (1) that forces are at work in each of them tending to produce normal distributions, and (2) that these forces are not independent of each other,[2] the probable error is computed in keeping with the theory of error typical of such distributions.[3] May the significance of correlation coefficients in time series be tested in the same manner? The answer must be sought in an analysis of how *completely if at all the foregoing assumptions hold for such* series.

As was noted above, time series are ordered in time, that is, each successive item holds its position in relation to the others, a succession of items of similar size tending to be the rule rather than the exception. In *non-time or condition* (attribute) series the order of the items has no significance. Moreover, in time series, random selection does not hold for the period of time for which trends, seasonal variations, and cyclical changes are determined. In fact a *specific* period is *selected by design,* care being taken to omit years which are exceptional—as for instance those during wars. The omission or inclusion of a year or of years may alter not only the trend but also the variations from trend for which characteristic pictures are being sought. The case is different with non-time series, the intent being to select at *random* as large a proportion of the population as is possible.

It is apparent, therefore, that probable errors computed for

[1] See the discussion of this subject by Professor Persons in the *Review of Economic Statistics,* April, 1919, pp. 124-127; "Correlation of Time Series" in *Handbook of Mathematical Statistics,* Houghton Mifflin, Boston, 1924, pp. 150-165 at pp. 162-163; "Some Fundamental Concepts of Statistics," *Journal of the American Statistical Association,* March, 1924, pp. 1-8, at pp. 6-8.

[2] See the discussion, pp. 406-410.

[3] See the discussion, pp. 428-429.

coefficients of correlation between time series, even though the latter are corrected for trend, and for seasonal and cyclical variations, do not have a probability meaning. As Professor Persons says:

"Thus, the 'probable error' of 0.03 in a coefficient of correlation of + 0.75 between the monthly items of pig-iron production and money rates six months later does not indicate, as one would conclude from the theory of probability, that the chances are billions to one against the independence of the two variables; *or, to state the idea more* specifically, that if we compute a coefficient from data of 'any' other actual period the chances are more than ten millions to one that its value would be over + 0.50. In fact, the significance of the 'probable error' of a constant computed from time series is not known, and, in practice, we do not view the world from the standpoint of mathematical probability. So that we are not surprised when we actually find that the coefficient of correlation between the adjusted figures for pig-iron production and money rates six months later for the period 1915-1918 is only + 0.38. We find sufficient explanation of this result, which is almost impossible and really astounding when viewed from the standpoint of random sampling, in the war demands for pig-iron, the tremendous imports of gold, government financing, *and the inauguration of the Federal reserve system during the period* in question. Neither are we surprised when we find that for the period 1919-1923 the maximum correlation between the two series is for a lag in money rates, not of six months, but of nine to twelve months. For this period includes the severe crisis and great financial stringency of 1920-1921, which dominated most of the items and hence the results. Thus in actual practice the statistician cannot reasonably assume ignorance of the peculiar circumstances pertaining to the special cases which constitute his material, and therefore he does not think in terms of random sampling and numerical probabilities. Granting as one must that consecutive items of a statistical time series are, in fact, related makes inapplicable the mathematical theory of probability."[1]

## VI. Conclusion

The treatment and correlation of time series involve the use of special statistical methods in many respects different

[1] Persons, W. M., "Some Fundamental Concepts of Statistics," *Journal of The American Statistical Association*, March, 1924, p. 7.

*from those commonly applied to other types of data.* These have to do with (1) the determination of long-time trends and short-time variations of different types; (2) their isolation; (3) the correction of original data for those influences; and (4) the correlation of the "corrected" series.

The technique of analysis, briefly described in this chapter, while developed for the most part in connection with the study of the business cycle, has general application wherever time series are involved. The importance to be attached to each of the steps, however, differs from problem to problem. The methods should not be applied blindly, nor should they always be considered superior to others, which from time to time have been and are being developed to suit special conditions. The end to be accomplished by analysis is always important, and the methods should be selected which will best help to realize it.

## References

*American Telephone and Telegraph Company*, "Statistical Analysis and Projection of Time Series," *Statistical Bulletin,* Number 4, Statistical Methods Series, Comptroller's Department (Issued by the Chief Statistician), New York, April, 1922.

"Comparison of an Individual Concern with the Harvard Index of General Business," *Harvard Economic Service,* Cambridge, Mass., 1924.

Crum, W. L., "Progressive Variation in Seasonality," *Journal of the American Statistical Association,* March, 1925, pp. 48-64.

Crum, W. L., & Patton, A. C., *Introduction to the Methods of Economic Statistics,* Part II, *The Analysis of Time Series,* The Harty-Husch Press, New Haven, Conn., pp. 65-82.

Davies, G. R., *Introduction to Economic Statistics,* Century Co., New York, 1922, pp. 100-148.

Falkner, H. D., "The Measurement of Seasonal Variation," *Journal of the American Statistical Association,* June, 1924, pp. 167-179. See References there given.

Hall, L. W., "A Moving Secular Trend and Moving Integration," *Journal of the American Statistical Association,* March, 1925, pp. 13-24.

KING, W. I., "An Improved Method for Measuring the Seasonal Factor," *Journal of the American Statistical Association*, September, 1924, pp. 301-313.

MILLS, FREDERICK C., *Statistical Methods Applied to Economics and Business*. Holt, New York, 1924, Chapters VII, pp. 252-314; VIII, pp. 315-343; XI, pp. 410-431.

PERSONS, W. M., "Construction of a Business Barometer Based Upon Annual Data," *American Economic Review*, December, 1916, pp. 739-769.

PERSONS, W. M., "A Non-technical Explanation of the Index of General Business Conditions," *The Review of Economic Statistics*, February, 1920, Harvard Committee on Economic Research, Cambridge, Mass.

PERSONS, W. M., "Correlation of Time Series," *Handbook of Mathematical Statistics*, Houghton Mifflin, Boston, 1924, pp. 151-165.

PERSONS, W. M., "Indices of General Business Conditions," *The Review of Economic Statistics*, January and April, 1919, *passim*, Harvard University Press, 1919. This is the most important single reference that can be given to the treatment of this subject.

PERSONS, W. M., "Some Fundamental Concepts of Statistics," *Journal of the American Statistical Association*, March, 1924, pp. 1-8.

PERSONS, FOSTER, and HETTINGER, Editors, *The Problem of Business Forecasting*, Houghton Mifflin, Boston, 1924. This volume is made up of the papers presented at the Eighty-fifth Annual Meeting of the American Statistical Association, Washington, D. C., December, 1923.

# CHAPTER XV

## THE PRINCIPLES OF INDEX NUMBER MAKING AND USING

### I. Introduction

Business men and students of economics and of social affairs use index numbers to measure changes in prices, wages, sales, production, stocks, and a multitude of other phenomena over a period of time. Rarely, however, are the sources of the data upon which they rest, the methods by which they are computed, and their suitability to special uses given consideration.

The fact that index numbers are supposed to measure changes in such elusive things as prices of commodities and services, for instance, differing at different times, in different markets, and under varying conditions of sale and methods of calculation ought to be sufficient warning against their hasty use. But, unfortunately, this is not the case. Those which are designed for some special purpose are given general application, while those which are intended to measure general changes are applied to specific uses with little or no thought of the consequences. Their use and preparation are too often divorced. This comes about because index numbers of a variety of types—not easily distinguished as to purpose, method of calculation, etc., by the layman—are easily obtained, and because those who have occasion to use index numbers rarely have the time and training to prepare them. Instruction in both index number making and using is needed. It is the purpose of this and the following chapter to furnish a basis for such instruction.

## II. Index Numbers Defined and the Methods of Computing Them Illustrated

Index numbers are a series of numbers by which changes in the magnitude of a phenomenon are measured from time to time or from place to place. For example, the number, 176, which shows the relation of the average wholesale price of a group of commodities in 1924 to their price in 1913 is an index number. The series of numbers expressing similar relations for prices in each year from 1913 to 1924 are known as index numbers. Moreover, the same expression is applied to numbers which show changes in prices between two or more places. Their purpose, therefore, is to reduce to a common denominator the qualities of different phenomena—as prices, stocks, production, etc.—so as to allow time and place comparisons to be made.

But

". . . it must be borne in mind that no index number corresponds to a real thing. It is not like the mean of certain observations in natural science—such, for example, as those for measuring the distance between the earth and the sun—of which any one may err, but whose average will point to a single specific fact. An index number points to no single fact. It gives, to repeat, only an indication of a general trend of prices. People often think and speak loosely on this topic, as if an index number told the whole story once for all. There is no one change in prices. There is a medley of many changes, different in direction and degree. All that we can hope to secure by averaging and summarizing is some concise statement of the general drift."[1]

The nature of an index number and the methods by which it may be computed may be illustrated by means of an example.

An index number is wanted which will show the movement of wholesale prices of paper in Chicago from 1913 to 1921. Price data are available from books of jobbers on the following types of paper: "newsprint," "wrapping," "book," "fine," "paper-board," and "miscellaneous." How can these different

[1] Taussig, F. W., *Principles of Economics* (Revised Edition, 1915), Macmillan, New York, Vol. I, p. 294.

phenomena—"prices" of different grades of "paper"—be reduced to a common denominator so as to allow a time comparison to be made?

The prices come from different jobbers, apply to different kinds of paper and are yearly averages. Accordingly, both prices and paper must be made comparable. The prices may be *averaged*, and quoted for uniform quantities—100 lbs. The types of paper to which they apply cannot be averaged, but can be compared for the different jobbers so as to secure uniform grades. Reserving for later discussion the principles which such a problem presents, various index numbers of prices may be constructed.

The average yearly prices and the types of paper used in the illustration are shown in Table 78.

TABLE 78

AVERAGE WHOLESALE PRICES OF DIFFERENT TYPES OF PAPER IN CHICAGO, 1913–1921

| LINE | TYPES OF PAPER | NUMBER OF GRADES | AVERAGE PRICES IN UNITS OF 100 LBS. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1913 | 1914 | 1915 | 1916 | 1917 | 1918 | 1919 | 1920 | 1921 |
| 1 | Newsprint * | 1 | $3.25 | $3.25 | $3.25 | $5.07 | $6.56 | $5.60 | $6.31 | $11.94 | $8.19 |
| 2 | Wrapping † | 2 | 4.53 | 4.27 | 4.24 | 7.52 | 9.90 | 9.92 | 9.56 | 14.56 | 10.53 |
| 3 | Book ‡ | 6 | 6.60 | 6.61 | 6.70 | 9.75 | 11.28 | 12.08 | 13.16 | 19.54 | 14.50 |
| 4 | Fine § | 11 | 10.81 | 10.90 | 11.29 | 15.38 | 17.98 | 19.93 | 22.85 | 29.51 | 24.49 |
| 5 | Paper-board ‖ | 4 | 4.75 | 4.75 | 4.73 | 6.42 | 7.73 | 8.72 | 9.58 | 12.55 | 9.72 |
| 6 | Miscellaneous ¶ | 8 | 9.12 | 9.19 | 9.49 | 13.99 | 16.97 | 18.66 | 20.85 | 27.26 | 23.30 |
| 7 | Average | — | 6.51 | 6.49 | 6.62 | 8.02 | 11.74 | 12.49 | 13.72 | 19.23 | 15.12 |

* Standard Newsprint.
† Kraft, Manila.
‡ Sized and super-calendered, Machine finished, Eggshell, Coated, Coated (high grade), Cover.
§ Ledger (cheap), Ledger (medium), Ledger (good), Bond (cheap), Bond (medium), Bond (good), Writing (manila), Writing (medium), Writing (good), Writing (French), Onion skin.
‖ Bristol, Straw, Jute, Pulp.
¶ Document manila, Blotting (white), Envelopes.

## 1. THE AVERAGE OF RELATIVES (RATIOS) METHOD

### (1) "Simple" Average of Relatives (Ratios)

#### a. Fixed Base

If the 1913 average price of each type of paper is taken as 100, and the price in each of the other years is expressed as a

percentage of this amount, and multiplied by 100, the relatives shown in Table 79, lines 1 to 6, are secured.

The process of computing the relatives or percentages may be illustrated as follows: The average price of newsprint in 1916 was \$5.07 per 100 lbs. The average price of the corresponding type of paper in the base year, 1913, was \$3.25. Accordingly, the relative price of this paper in 1916 was $\frac{\$5.07}{\$3.25} \times 100 = 156$. This number is a per cent, or as is indicated above, a relative. Similarly, the average price of "miscellaneous" paper in 1920 was \$27.26. The 1913 average price was \$9.12. Therefore, the relative price in 1920 was $\frac{\$27.26}{\$9.12} \times 100 = 299$. All of the relatives in Table 79 are computed in this manner.

TABLE 79

RELATIVE WHOLESALE PRICES OF PAPER IN CHICAGO
1913 TO 1921
(1913 = 100)

| Line | Types of Paper | Percentages or Relatives—1913 = 100 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1913 | 1914 | 1915 | 1916 | 1917 | 1918 | 1919 | 1920 | 1921 |
| 1 | Newsprint ...... | 100 | 100 | 100 | 156 | 202 | 172 | 194 | 367 | 252 |
| 2 | Wrapping ...... | 100 | 94 | 94 | 166 | 219 | 219 | 211 | 321 | 232 |
| 3 | Book .......... | 100 | 100 | 102 | 148 | 171 | 183 | 199 | 296 | 220 |
| 4 | Fine ........... | 100 | 101 | 104 | 142 | 166 | 184 | 211 | 273 | 227 |
| 5 | Paper-board .... | 100 | 100 | 100 | 135 | 163 | 184 | 202 | 264 | 205 |
| 6 | Miscellaneous ... | 100 | 101 | 104 | 153 | 186 | 205 | 229 | 299 | 255 |
| 7 | Total of Relatives ....... | 600 | 596 | 604 | 900 | 107 | 1147 | 1246 | 1820 | 1391 |
| 8 | Average of Relatives ... | 100 | 99 | 101 | 150 | 185 | 191 | 208 | 303 | 232 |
| 9 | Median ....... | 100 | 100 | 101 | 151 | 179 | 184 | 207 | 298 | 230 |
| 10 | Geometric Mean. | 100 | 99 | 101 | 150 | 183 | 191 | 207 | 303 | 231 |

Lines 8, 9, and 10, respectively, of Table 79 show arithmetic means, medians, and geometric means computed from these relatives.

The arithmetic mean in each year is the result of dividing the sum of the relatives by six. The medians are secured by arranging the relatives each year in order of magnitude and taking the middle item. In all but three years—1913, 1914, and 1918—interpolation was necessary in order to find a precise median.[1]

The geometric mean of relatives each year is gotten by multiplying together the relatives and taking the 6th root. This is done by logarithms as follows: (1) find the log of each of the relatives, (2) add the logs together, (3) divide the sum by 6, and (4) look up the natural number corresponding to the product in (3). The natural number is the index for the year in question.

### b. Chain Base

In Table 79 the relative or percentage numbers are based on 1913. In Table 80, however, they are based on the preceding year. That is, the years are linked together. Line 7 gives the averages of the link-relatives, and line 8, the chain-relatives based on 1913.

The chain-relatives are secured from the average link-relatives as follows: The average link-relative for 1913—100—is multiplied by the link-relative for 1914 on 1913—99. This gives the chain-relative, 99, for 1914 on 1913. The chain-relative for 1915 on 1913—100—is secured by multiplying the link-relative for 1914 on 1913—99—by the link-relative for 1915 on 1914—101. The chain-relative for 1916 on 1913—150—is secured by multiplying the link-relatives—99 × 101 × 150. The remaining chain-relatives are secured in a similar manner.

[1] See Chapter IX, pp. 286-289, for a discussion of interpolation for medians.

The amounts in line 8 are chain index numbers based upon 1913. Those in line 7 are relative or percentage numbers showing average year-to-year changes.

TABLE 80

TABLE SHOWING CHAIN-RELATIVE INDEX NUMBERS OF WHOLESALE PRICES OF PAPER IN CHICAGO, 1913 TO 1921
(1913 = 100)

| Line | Types of Paper | Percentages or Relatives Based on Preceding Year | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1913 | 1914 | 1915 | 1916 | 1917 | 1918 | 1919 | 1920 | 1921 |
| 1 | Newsprint ........ | *100* | *100* | *100* | *156* | *129* | *85* | *113* | *189* | *69* |
| 2 | Wrapping ......... | 100 | 94 | 99 | 177 | 132 | 100 | 96 | 152 | 72 |
| 3 | Book ............. | 100 | 100 | 101 | 146 | 116 | 107 | 109 | 149 | 74 |
| 4 | Fine ............. | 100 | 101 | 104 | 136 | 117 | 111 | 115 | 129 | 83 |
| 5 | Paper-board ....... | 100 | 100 | 100 | 136 | 120 | 113 | 110 | 131 | 77 |
| 6 | Miscellaneous ..... | 100 | 101 | 103 | 147 | 121 | 110 | 112 | 131 | 85 |
| 7 | Average Link-Relatives ....... | 100 | 99 | 101 | 150 | 123 | 104 | 109 | 147 | 77 |
| 8 | Chain-Relatives 1913 = 100 ..... | 100 | 99 | 100 | 150 | 185 | 192 | 209 | 307 | 236 |

### (2) *Weighted Average of Relatives (Ratios)*

In Table 79, the relative price of each type of paper is counted once in order to secure the index based on averages of relatives—a so-called unweighted figure. That is, the sum of the relatives in each year is divided by six. If weights, proportional to the value of each type of paper consumed in the United States, are assigned to the relatives, the *weighted* average of relatives index is as given in Table 81—line 8.[1]

[1] Neither the quantity nor the value of these types of paper consumed in Chicago is available. Quantity weights for the United States in 1917 are found in Mitchell, W. C., *History of Prices During the War*, Bulletin No. 31, Averill, W. A., "Prices of Paper," *War Industries Board*, Washington, D. C., 1919. They are given on a proportional basis in Table 85. For a weighted average of relatives index number, however, *value* weights are desired. They may be secured from the quantity weights in Table 85 as follows: (1) compute a weighted average price of all grades of paper by multiplying the average value, type by type in Table 78, by the corresponding weights as shown in Table 85—the average value is

TABLE 81

TABLE GIVING WEIGHTED AVERAGE OF RELATIVES INDEX NUMBERS OF WHOLESALE PRICES OF PAPER IN CHICAGO, 1913 TO 1921. BASE WEIGHTS: VALUE OF PAPER CONSUMED IN 1917

(1913 = 100)

| LINE | TYPE OF PAPER | VALUE WEIGHT; PER CENT OF TOTAL CONSUMPTION, 1917 | PRODUCTS OF WEIGHTS AND RELATIVES (SEE TABLE 79) TO NEAREST WHOLE NUMBER | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1913 | 1914 | 1915 | 1916 | 1917 | 1918 | 1919 | 1920 | 1921 |
| 1 | Newsprint | 20.4 | 2,040 | 2,040 | 2,040 | 3,182 | 4,121 | 3,509 | 3,958 | 7,487 | 5,141 |
| 2 | Wrapping | 12.4 | 1,240 | 1,166 | 1,166 | 2,058 | 2,716 | 2,716 | 2,616 | 3,980 | 2,877 |
| 3 | Book | 18.4 | 1,840 | 1,840 | 1,877 | 2,723 | 3,146 | 3,367 | 3,662 | 5,446 | 4,048 |
| 4 | Fine | 14.3 | 1,430 | 1,444 | 1,487 | 2,031 | 2,374 | 2,631 | 3,017 | 3,904 | 3,246 |
| 5 | Paper-board | 27.3 | 2,730 | 2,730 | 2,730 | 3,686 | 4,450 | 5,023 | 5,515 | 7,207 | 5,597 |
| 6 | Miscellaneous | 7.2 | 720 | 727 | 749 | 1,102 | 1,339 | 1,476 | 1,649 | 2,153 | 1,836 |
| 7 | Total | 100.0 | 10,000 | 9,947 | 10,049 | 14,782 | 18,146 | 18,722 | 20,417 | 30,177 | 22,745 |
| 8 | Weighted Average * | | 100 | 99 | 100 | 148 | 181 | 187 | 204 | 302 | 227 |

* Products in Line 7 divided by sum of the weights, 100.

In order to secure yearly index numbers, the relative for each type of paper each year is multiplied by the value weight in 1917, the products totaled, and divided by the sum of the weights, 100. *For example, the relative for newsprint in* 1916 based on 1913 is 156. The value weight for this type of paper in 1917 is 20.4. Accordingly, the product of the relative and the weight, $156 \times 20.4$, is 3182. The corresponding product for wrapping paper is 2058; for book paper, 2723. The products *for the other types in this year are given in the column*

Note 1 continued

$5.08 per 100 lbs.; (2) express as a proportion of this quantity the *average value of each type secured by multiplying the average price by a* percentage representing its portion of the total quantity. For example: the average price of newsprint in 1913 was $3.25. Newsprint was 32 per cent of the total consumed. Therefore, 32 per cent of $3.25 = $1.04, which is 20.5 per cent of $5.08, the weighted average value. The weights for the other types are computed in the same manner.

If the prices of the different types of paper were expressed in different units—as, for instance, in 100 lbs., in rolls, in tons, etc.—it would be necessary to use weights measured in corresponding units. In this case, however, since the units are the same, the weights may be put on a proportional basis.

for 1916. The sum of the weights is 100; therefore, the weighted average of relatives index number for 1916 is $\frac{14{,}782}{100} = 148$.

The series of amounts in line 8 are weighted averages of relatives index numbers based on 1913.

An index number based upon weighted medians of relatives is shown in Table 82, the weights for the different types of paper being the estimated proportions of the value consumed. In order to calculate weighted medians, the relatives must be arranged in order of magnitude, and the corresponding weights accompany them. The weights are the frequencies which must be divided into two equal parts in order to calculate the medians.[1]

TABLE 82

TABLE SHOWING WEIGHTED MEDIANS OF RELATIVES INDEX NUMBERS OF WHOLESALE PAPER PRICES, CHICAGO, 1913 TO 1921 (1913 = 100)

| YEAR | INDEX NUMBER |
|---|---|
| 1913 | 100 |
| 1914 | 100 |
| 1915 | 100 |
| 1916 | 148 |
| 1917 | 171 |
| 1918 | 184 |
| 1919 | 202 |
| 1920 | 296 |
| 1921 | 227 |

Table 83 illustrates the manner in which the relatives and the weights (frequencies) must be arranged in order to find the median. The arrangement refers to 1916. It should be observed that the order may and probably will be different each year.

[1] See formulæ for medians, *supra*, p. 283.

TABLE 83

TABLE SHOWING THE METHOD OF COMPUTING A WEIGHTED MEDIAN OF RELATIVES INDEX NUMBER OF WHOLESALE PAPER PRICES IN CHICAGO, 1916

| TYPES OF PAPER | RELATIVES 1916 Base = 1913 | VALUE WEIGHTS 1917 Per Cent |
|---|---|---|
| Paper-board | 135 | 27.3 |
| Fine | 142 | 14.3 |
| Book | 148 | 18.4 |
| Miscellaneous | 153 | 7.2 |
| Newsprint | 156 | 20.4 |
| Wrapping | 166 | 12.4 |
| Total | .... | 100 |
| Weighted Median of Relatives | 148 | .... |

## 2. RATIOS OF AVERAGES

An alternative method to averaging the relatives (ratios) unweighted (see Table 79) or weighted (see Table 81) is to express the average price each year in the form of a ratio relative to the price in a base year.

In Table 78, the prices for the different types of paper each year are given in units of 100 lbs. Line 7 of this table shows the simple average price in each of the years. If the different averages in this line are expressed as ratios with 1913 as a base, the index numbers are as given in Table 84.

That is, the average price in 1913, $6.51, is taken as 100, the average prices in the other years being expressed as percentages of this amount and multiplied by 100. For instance, the index number for 1917 is $\frac{\$11.74}{\$6.51} \times 100 = 180$. The index numbers for the other years are computed in a similar manner.

Either the average price or the sum of the average prices

may be expressed in this manner. Since the totals are divided by the same amount—six in this case—in order to get the averages, the relations between the ratios for the different years are identical in the two methods.

TABLE 84

TABLE SHOWING RATIOS-OF-AVERAGES INDEX NUMBERS OF WHOLESALE PAPER PRICES IN CHICAGO, 1913-1921
(1913 = 100)

| YEARS | AVERAGE PRICE | INDEX NUMBER 1913 = 100 |
|---|---|---|
| 1913 | $ 6.51 | 100 |
| 1914 | 6.49 | 100 |
| 1915 | 6.62 | 102 |
| 1916 | 8.02 | 123 |
| 1917 | 11.74 | 180 |
| 1918 | 12.49 | 192 |
| 1919 | 13.72 | 211 |
| 1920 | 19.23 | 295 |
| 1921 | 15.12 | 232 |

### 3. RATIOS OF WEIGHTED AGGREGATES

Instead of using (1) different unweighted averages of relatives (as in Table 79), (2) different weighted averages of relatives (as in Tables 81 and 82), or (3) ratios of averages (as in Table 84), the actual prices may be weighted by suitable quantities, totaled or aggregated, and expressed as ratios relative to a given base. Index numbers computed in this manner are given in Table 85, 1913 being used as the base.

The method of computing this type of an index is different from that used in Table 81. In Table 85, the actual prices are weighted by quantities; in Table 81, the relative prices are weighted by values. It will be noticed, however, that the results are the same.[1] The reason for this agreement is well

[1] A slight difference occurs for the year 1914, but this is due to the treatment of decimal amounts.

TABLE 85

TABLE GIVING WEIGHTED AGGREGATE OF ACTUAL PRICES INDEX NUMBERS OF WHOLESALE PRICES OF PAPER IN CHICAGO, 1913 TO 1921

BASE WEIGHTS AS PROPORTIONS CONSUMED IN 1917 (1913 = 100)

| LINE | TYPES OF PAPER | QUANTITY WEIGHTS: PER CENTS OF TOTAL CONSUMPTION, 1917 | PRODUCTS OF PRICE (SEE TABLE 78) AND WEIGHTS (SEE COLUMN 3)—PER CENTS OF TOTAL CONSUMPTION IN THE UNITED STATES | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1913 | 1914 | 1915 | 1916 | 1917 | 1918 | 1919 | 1920 | 1921 |
| 1 | Newsprint | 32.0 | 104.0 | 104.0 | 104.0 | 162.2 | 209.9 | 179.2 | 201.9 | 382.1 | 262.1 |
| 2 | Wrapping | 13.9 | 63.0 | 59.4 | 58.9 | 104.5 | 137.6 | 137.9 | 182.9 | 202.4 | 146.4 |
| 3 | Book | 14.2 | 93.7 | 93.9 | 95.1 | 138.5 | 160.2 | 171.5 | 186.9 | 277.5 | 205.9 |
| 4 | Fine | 6.7 | 72.4 | 73.0 | 75.6 | 103.0 | 120.5 | 133.5 | 153.1 | 197.7 | 164.1 |
| 5 | Paper-board | 29.2 | 138.7 | 138.7 | 138.1 | 187.5 | 225.7 | 254.6 | 279.7 | 366.5 | 283.8 |
| 6 | Miscellaneous | 4.0 | 36.5 | 36.8 | 38.0 | 56.0 | 67.9 | 74.6 | 83.4 | 109.0 | 93.2 |
| 7 | Total | | 598.3 | 505.8 | 509.7 | 751.7 | 921.8 | 951.3 | 1037.9 | 1535.2 | 1155.5 |
| 8 | Relatives * 1913 = 100 | | 100 | 100 | 100 | 148 | 181 | 187 | 204 | 302 | 227 |

* To the nearest whole number.

expressed by Mitchell. He says:

"... if we want an aggregate of actual prices, we merely multiply the quotations of each commodity at each date by the physical quantities used as weights, and add these products. To measure the variations of these aggregates in terms of prices at the base period, we have only to divide the aggregate for each period by the aggregate for the base period. But if we plan to make a weighted arithmetic mean of price variations, we begin by turning the quotations into relative prices. That is, we divide the actual price of each commodity at each date by its price in the base period. Then we weight these relatives, not by physical quantities as in the first case, but by the money values of the physical quantities at the prices of the base year. But in this step the prices of the base year, which were just used as divisors to get relative prices, are used again as factors by which the relative prices are multiplied. Hence our results are the same as if we had neither multiplied nor divided by the prices of the base year; in other words, the same as if we had multiplied the quotations of each commodity in each year by the physical quantities used as weights. But that is just what we did when we set out to make an aggregate of actual prices. So far, then, the two processes are identical in their outcome. And the remaining steps are also the same.

The products must be added, and the sums divided by the physical quantities used as weights times the actual prices of the base year. Therefore, to make relative prices from aggregates of actual prices is a shorter way of getting the same results as are obtained by making similarly weighted arithmetic means of relative prices."[1]

### 4. SUMMARY OF RESULTS BY DIFFERENT METHODS

Different methods of computing index numbers for the wholesale prices of six types of paper in Chicago give varying results. These are compared in Table 86.

TABLE 86

INDEX NUMBERS OF WHOLESALE PRICES OF PAPER IN CHICAGO 1913-1921 COMPUTED BY DIFFERENT METHODS
(1913 = 100)

| Year | Averages of Relatives (Ratios) | | | | | | Ratio of Averages | Weighted Aggregate of Actual Prices * |
|---|---|---|---|---|---|---|---|---|
| | Unweighted | | | | Weighted | | | |
| | Arithmetic Mean | | Median | Geometric Mean | Arithmetic Mean * | Median | | |
| | Fixed Base | Chain Base | | | | | | |
| 1913 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 1914 | 99 | 99 | 100 | 99 | 99 | 100 | 100 | 100 |
| 1915 | 101 | 100 | 101 | 101 | 100 | 100 | 102 | 100 |
| 1916 | 150 | 150 | 151 | 150 | 148 | 148 | 123 | 148 |
| 1917 | 185 | 185 | 179 | 183 | 181 | 171 | 180 | 181 |
| 1918 | 191 | 192 | 184 | 191 | 187 | 184 | 192 | 187 |
| 1919 | 208 | 209 | 207 | 207 | 204 | 202 | 211 | 204 |
| 1920 | 303 | 307 | 298 | 303 | 302 | 296 | 295 | 302 |
| 1921 | 232 | 236 | 230 | 231 | 227 | 227 | 232 | 227 |

* See the comment, p. 478, relative to the results by these methods.

In some cases the differences are large, in others, negligible. For two methods the numbers are identical throughout. Moreover, for certain years all methods give the same results.

[1] Mitchell, *op. cit.*, 80-81.

This single body of price data has served to illustrate the arithmetic of the more important methods of computing index numbers. The remaining discussion of the chapter is concerned with the principles back of the methods. It will help to explain the reasons for the differences and similarities.

## III. The Uses of Index Numbers

In what has gone before, plan and purpose in statistical study have been emphasized. Both need to be especially stressed in connection with index numbers, because, while most of those that are currently used are of the "general purpose" type, they are given a variety of special uses.

"Few of the widely used index numbers, . . . are made to serve one special purpose. On the contrary, most of them are 'general-purpose' series, designed with no aim more definite than that of measuring changes in the price level. Once published they are used for many ends—to show the depreciation of gold, the rise in the cost of living, the alternations of business prosperity and depression, and the allowance to be made for changed prices in comparing estimates of national wealth or private income at different times. They are cited to prove that wages ought to be advanced or kept stable; that railway rates ought to be raised or lowered; that 'trusts' have manipulated the prices of their products to the benefit or the injury of the public; that tariff changes have helped or harmed producers or consumers; that immigration ought to be encouraged or restricted; that the monetary system ought to be reformed; that natural resources are being depleted or that the national dividend is growing. They are called in to explain why bonds have fallen in price and why interest rates have risen, why public expenditures have increased, why social unrest prevails in certain years, why farmers are prosperous or the reverse, why unemployment fluctuates, why gold is being imported or exported, and why political 'landslides' come when they do."[1]

Generally speaking, however, two major purposes, so far as price indexes are concerned, are distinguishable: (1) to meas-

[1] Mitchell, Wesley C., "Index Numbers of Wholesale Prices in the United States and Foreign Countries," *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 173, July, 1915, pp. 25-26.

ure general changes in prices, and (2) to interpret the effect of the changes upon various classes of people.

An index number serving the first use is computed from the prices of a wide selection of commodities covering all phases of industry; one designed for the second purpose, from the commodities the changes in prices of which have special reference to the class concerned. For instance, the United States Bureau of Labor Statistics publishes index numbers of wholesale prices based upon 404 commodities, the selection being made with the intent of sampling the general market.[1] On the other hand, the same Bureau publishes index numbers of retail prices of foods, the commodities being selected from industrial centers and referring to articles currently purchased by so-called workingmen's families.[2] Their purpose is to serve as a basis for approximating the effect of price changes upon consumers. A variety of special purpose types of index numbers are now issued, the more important of which are described in Chapter XVI.

But index numbers are not restricted to price phenomena. Any phenomenon extending over a period of time and expressed numerically may be put in this form, the only peculiarity being that its relative rather than its absolute aspect is exhibited. Index numbers of wages, rents, imports, exports, sales, production, or of any other phenomenon may be constructed. Some of the more important of these non-price series are described in Chapter XVI.

## IV. Principles of Index Number Making

Because the uses which are made of index numbers of prices and of other phenomena vary widely, and because different methods are available according to which they may be constructed, the question of the purpose which they are to serve is of first importance.

[1] See the discussion of this index number, Chapter XVI, pp. 516-518.
[2] See the discussion of this index number, Chapter XVI, pp. 520-521.

Generally speaking, the purpose of an index number is, as Fisher says, "that it shall *fairly represent,* so far as one single figure can, the general trend of the many diverging ratios from which it is calculated. It should be a 'just compromise' among conflicting elements, the 'fair average,' the 'golden mean.' Without some kind of fair splitting of the differences involved, an index number is apt to be unsatisfactory, if not absurd." [1] The difficulty of securing such a "fair average" can be appreciated only by a detailed study of the index numbers currently issued, and of the principles involved in index number making.[2]

## 1. THE ATTRIBUTES OF INDEX NUMBERS AND THE STEPS IN THEIR CONSTRUCTION

Fisher enumerates as follows the attributes of an index number:

*(1) "As to the Construction of the Index Number*

a. *"The general character of the data included, e.g.* 'wholesale prices' or 'retail prices' of commodities, or 'prices of stocks,' or 'wages,' or 'volume of production,' etc.

b. *"The specific character of data included, e.g.* 'foods,' still further specified as 'butter,' 'beef,' etc.

c. *"Their assortment, e.g.* a larger proportion of quotations of meats than of vegetables.

d. *"The number of quotations used, e.g.* '22 commodities' as in the case of the *Economist* index number (until recently) as contrasted with '1474 commodities' as in the case of the War Industries Board.

[1] Fisher, Irving, *The Making of Index Numbers*, Houghton Mifflin, Boston, 1922, p. 10.

[2] Such a comparative study has been made by Professor Wesley C. Mitchell in "Index Numbers of Wholesale Prices in the United States and Foreign Countries," *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 284, October, 1921. Acknowledgments are here made of the indebtedness of the writer to Professor Mitchell for much of the illustrative matter in this and the following chapters. An elaborate analysis of a somewhat different kind has also been made by Professor Fisher in his monumental study, *The Making of Index Numbers*, referred to immediately above.

e. "*The kind of mathematical formula* employed for calculating the index number, *e.g.* the 'simple arithmetic average' or the 'weighted geometric average,' etc.

*(2) "As to the Particular Times or Places to Which the Index Number Applies*

a. "*The period covered, e.g.* '1913-1918,' or the territory covered, *e.g.* certain specified cities of which the price levels are to be compared.

b. "*The base, e.g.* the year 1913.

c. "*The interval between successive indexes, e.g.* 'yearly' or 'monthly.'

*(3) "As to the Sources and Authorities*

a. "*The agency which collects, calculates, and publishes the index number, e.g.* 'Bradstreet's' or the 'United States Bureau of Labor Statistics.'

b. "*The markets used, e.g.* the 'Stock' or 'Produce' Exchanges of 'New York' or the 'primary markets of the United States.'

c. "*The sources of quotations, e.g.* the 'leading trade journals' or the books of business houses.

d. "*The publications containing the index number, e.g.* the Bulletin of the United States Bureau of Labor Statistics."[1]

Mitchell approaches the problem somewhat differently. His enumeration of the processes in making an index number is as follows:

"(1) Defining the purpose for which the final results are to be used; (2) deciding the numbers and kinds of commodities to be included; (3) determining whether these commodities shall all be treated alike or whether they shall be 'weighted' according to their relative importance; (4) collecting the actual prices of the commodities chosen, and, in case a weighted series is to be made, collecting also data regarding their relative importance; (5) deciding whether the form of the index number shall be one showing the average variations of prices or the variations of a sum of actual prices; (6) in case average variations are to be shown, choosing the base upon which relative prices shall be computed; and (7) settling upon the form of average to be struck, if averages are to be used.

[1] Fisher, Irving, *The Making of Index Numbers*, Houghton Mifflin, Boston, 1922, pp. 8-9.

"At each one of these successive steps choice must be made among alternatives that range in number from two to thousands. The possible combinations among the alternatives chosen are indefinitely numerous. Hence there is no assignable limit to the possible varieties of index numbers, and in practice no two of the known series *are exactly alike in construction. To canvass even the important* variations of method actually in use is not a simple task."[1]

## 2. DATA FROM WHICH PRICE INDEX NUMBERS ARE MADE

In a study of prices attention must first be centered upon the commodities included and the conditions of price making. Distinction will have to be made between producers' and consumers' goods,[2] between raw and manufactured commodities,[3] between manufactured goods bought by consumers for family

[1] Mitchell, Wesley C., "Index Numbers of Wholesale Prices in the United States and Foreign Countries," *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 284, October, 1921, p. 23.

[2] ". . . there are characteristic differences between the price fluctuations of manufactured commodities bought by consumers for family use and the price fluctuations of manufactured commodities bought by business *men for industrial or commercial use. . . . Though consisting more* largely of the erratically fluctuating farm products, the consumers' goods *are steadier in price than the producers' goods, because the demand for* them is less influenced by changes in business conditions." *Op. cit.*, pp. 46-48.

[3] "These several comparisons establish the conclusion that manufactured goods are steadier in price than raw materials. The manufactured goods fell less in 1890-1896, rose less in 1896-1907, again fell less in 1907-1908, and rose less in 1908-1913. Further, the manufactured goods had the narrower extreme range of fluctuations, the smaller average change from year to year, and the slighter advance in price from one decade to the next. It follows that index numbers made from the prices of raw materials, or of raw materials and slightly manufactured products, must be expected to show wider oscillations than index numbers including a liberal representation of finished commodities." *Op. cit.*, p. 41.

"First, the list of commodities used by the Bureau of Labor Statistics includes 29 quotations for iron and its products, 30 quotations for cotton and its products, and 18 for wool and its products, besides 8 more quotations for fabrics made of wool and cotton together. On the other hand it has but 7 series for wheat and its products, 8 for coal and its products, 3 for copper and its products, etc. The iron, cotton, and wool groups together make up 85 series out of 242, or 35 per cent of the whole number. . . .

"Does this large representation of three staples distort these index numbers—particularly the bureau's series where the disproportion is greatest? Perhaps, but if so the distortion does not arise chiefly from the undue influence assigned to the price fluctuations of raw cotton, raw wool, and pig iron. For, contrary to the prevailing impression, the similarity between the price fluctuations of finished products and their raw

use and manufactured commodities bought by business men for industrial uses,[1] between mineral products, animal products and farm crops,[2] etc., the prices of all of which respond dif-

*Note 3, continued.*

materials is less than the similarity between the price fluctuations of finished products made from different materials. . . . As babies from different families are more like one another than they are like their respective parents, so here the relative prices of cotton textiles, woolen textiles, steel tools, bread, and shoes differ far less among themselves than they differ severally from the relative prices of raw cotton, raw wool, pig iron, wheat, and hides. Hence the inclusion of a large number of articles made from iron, cotton, and wool affects an index number mainly by increasing the representation allotted to manufactured goods. What materials those manufactured goods are made from makes less difference in the index number than the fact that they are manufactured. To replace iron, cotton, and woolen products by copper, linen, and rubber products would change the results somewhat, but a much greater change would come from replacing the manufactured forms of iron, cotton, and wool by new varieties of their raw forms." *Op. cit.*, pp. 48-50.

[1] "It has been found that among manufactured commodities those bought for family consumption are steadier in price than those bought for business use." *Op. cit.*, p. 51.

[2] "Third, there are characteristic differences among the price fluctuations of the groups consisting of mineral products, forest products, animal products, and farm crops. . . . Fifty-seven commodities are included, all of them raw materials or slightly manufactured products. Here the striking feature is the capricious behavior of the prices of farm crops under the influence of good and bad harvests. The sudden upward jump in their prices in 1891, despite the depressed condition of business, their advance in the dull year 1904, their fall in the year of revival 1905, their failure to advance in the midst of the prosperity of 1906, their trifling decline during the great depression of 1908, and their sharp rise in the face of reaction in 1911 are all opposed to the general trend of other prices. The prices of animal products are distinctly less affected by weather than the prices of vegetable crops, but even they behave queerly at times, for example in 1893. Forest-product prices are notable chiefly for maintaining a much higher level of fluctuation in 1902-1913 than any of the other groups, a level on which their fluctuations, when computed as percentages of the much lower prices of 1890-1899, appear extremely violent. Finally, the prices of minerals accord better with alternations of prosperity, crisis, and depression than any of the other groups. And the anomalies that do appear—the slight rise in three years (1896, 1903, and 1913) when the tide of business was receding—would be removed if the figures were compiled by months. For the trend of mineral prices was downward in these years, but the fall was not so rapid as the rise had been in the preceding years, so that the annual averages were left somewhat higher than before. An index number composed largely of quotations for annual crops, then, would be expected at irregular intervals to contradict capriciously the evidence of index numbers in which most of the articles were mineral, forest, or even animal products." *Op. cit.*, pp. 44-46.

ferently to conditions of scarcity and surplus.[1] Obviously, a price index number which reflects price changes at large must be made from samples of all commodity groups that are affected *in a peculiar manner. Similarly, in using an index* number prepared by another, one must satisfy himself respecting the list of commodities used before he can be sure what in reality the index measures.

But what is meant by "price"? Has one in mind retail or wholesale price? price at what place? under what condition of sale? to whom? price of what grade of commodity? on what market? Are the "prices" contract, import, or market prices? What is *the* wholesale or retail price of a commodity?

"We commonly speak of *the* wholesale price of articles like pig iron, cotton, or beef as if there were only one unambiguous price for any one thing on a given day, however this price may vary from one day to another. In fact there are many different prices for every great staple on every day it is dealt in, and most of these differences are of the sort that tend to maintain themselves even when markets are highly organized and competition is keen. Of course varying grades command varying prices, and so as a rule do large lots and small lots; for the same grade in the same quantities, different prices are paid by the manufacturer, jobber, and local buyer; in different localities the prices paid by these various dealers are not the same; even in the same locality different dealers of the same class do not all pay the same price to every one from whom they buy the same grade in the same quantity on the same day. To find what really was the price of cotton, for example, on February 1, 1920, would require an elaborate investigation, and would result in showing a multitude of different prices covering a considerable range.

"Now the field worker collecting data for an index number must select from among all these different prices for each of his commodities the one or the few series of quotations that make the most representative sample of the whole. He must find the most reliable source of information, the most representative market, the most typical brands or grades, and the class of dealers who stand in the most influential position. He must have sufficient technical knowledge to be sure that his quotations are for uniform qualities, or to make the necessary adjustments if changes in quality have

[1] This topic has been given elaborate treatment by Professor Mitchell in his *Business Cycles* (University of California, Memoirs, Vol. III, September, 1913), pp. 93-109.

occurred in the markets and require recognition in the statistical office. He must be able to recognize anything suspicious in the data offered him and to get at the facts. He must know how commodities are made and must seek comparable information concerning the prices of raw materials and their manufactured products, concerning articles that are substituted for one another, used in connection with one another, or turned out as joint products of the same process. He must guard against the pitfalls of cash discounts, premiums, rebates, deferred payments, and allowances of all sorts. And he must know whether his quotations for different articles are all on the same basis, or whether concealed factors must be allowed for in comparing the prices of different articles on a given date."[1]

If it is difficult to establish *the* price of a commodity at *one* time it is even more difficult to guarantee that *the* price determined at one time is *the* price at some *other* time. Conditions of marketing change, commodities change as to quality and salability, and price lists of identical commodities for any great length of time are frequently not available. The paucity of price data and the unwillingness of people to place any reliance in those extant were undoubtedly the main reasons for the relatively late development of index numbers.[2]

Today, of course, such data as those from which the index numbers currently published by the United States Bureau of Labor Statistics are computed, are furnished by reputable firms and corporations, according to uniform instructions, on uniform blanks, and are carefully scrutinized by the agents of the Government.

But how many commodities are necessary in order that an index number may indicate either the amount or effect of price change? From what regions should prices be drawn, and how frequently ought they to be recorded? Are prices quoted in standard and definite units?[3] Some commodities

[1] *Op. cit.*, pp. 25-26.

[2] *Op. cit.*, p. 10.

[3] "Often the form of quotation makes all the difference between a substantially uniform and a highly variable commodity. For example, prices of cattle and hogs are more significant than prices of horses and mules, because the prices of cattle and hogs are quoted per pound, while the prices of horses and mules are quoted per head." *Op. cit.*, p. 33.

are sensitive to conditions of demand and supply; others react slowly under changed conditions. Some are vitally affected by seasons, while others show appreciable change only in the face of violent disturbance and exhibit a steady rise or fall only over long periods. "Typical" price behavior can hardly be predicted for any commodity. It may never occur.

What principles have been followed in the choice of commodities? Are raw and manufactured commodities disproportioned? Is a certain commodity unimportant for one purpose—or important for another—represented in both its raw and its manufactured state? How is the importance of a commodity given weight? What test of importance is applied? How is it measured? These are important questions which one must answer for himself for every index number before he uses it for a particular purpose.[1]

"Difficult as it is to secure satisfactory price quotations, it is still more difficult to secure satisfactory statistics concerning the relative importance of the various commodities quoted. What is wanted is an accurate census of the quantities of the important staples, at least, that are annually produced, exchanged, or consumed. To take such a census is altogether beyond the power of the private investigators or even of the Government bureaus now engaged in making index numbers. Hence the compilers are forced to confine themselves for the most part to extracting such information as they can from statistics already gathered by other hands and for other purposes than theirs. In the United States, for example, estimates of production, consumption, or exchange come from most miscellaneous sources: The Department of Agriculture, the Census Office, the Treasury Department, the Bureau of Mines, the Geological Survey, the Internal Revenue Office, the Mint, associations of manufacturers or dealers, trade papers, produce exchanges, traffic records of canals and railways, etc. The man who assembles and compares estimates made by these various organizations finds among them many glaring discrepancies for which it is difficult to account. Such conflict of evidence when two or more independent estimates of

[1] Both for American and European index numbers such questions as these and many more are answered in *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 284, to which reference has so frequently been made.

the same quantity are available throws doubt also upon the seemingly plausible figures coming from a single source for other articles. To extract acceptable results from this mass of heterogeneous data requires intimate familiarity with the statistical methods by which they were made, endless patience, and critical judgment of a high order, not to speak of tactful diplomacy in dealing with the authorities whose figures are questioned."[1]

Mitchell, following an elaborate comparison of the various American index numbers, so far as choice of commodities and the importance assigned them are concerned, arrives at the following conclusions:

"As for the small series made from the prices of foods alone or from the prices of any single group of commodities, it is clear that however good for special uses they may be, they are untrustworthy as general-purpose index numbers."[2]

"Large index numbers are more trustworthy for general purposes than small ones, not only in so far as they include more groups of related prices, but also in so far as they contain more numerous samples from each group. What is characteristic in the behavior of the prices of farm crops, of mineral products, of manufactured wares, of consumers' goods, etc.—what is characteristic in the behavior of any group of prices—is more likely to be brought out and to exercise its due effects upon the final results when the group is represented by 10 or 20 sets of quotations than when it is represented by only one or two sets. The basis of this contention is simple: In every group that has been studied there are certain commodities whose prices seldom behave in the typical way, and no commodities whose prices can be trusted always to behave typically. Consequently, no care to include commodities belonging to all the important groups can guarantee accurate results, unless care is also taken to get numerous representatives of each group."[3]

### 3. DISPERSION OF PRICE FLUCTUATIONS[4]

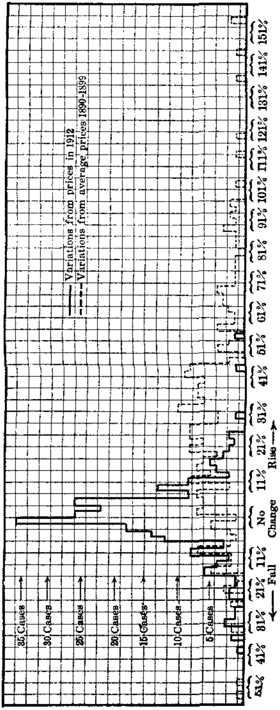
The trend of price change is generally in one direction for a considerable period. There are periods of falling and of

[1] *Op. cit.*, p. 26.
[2] *Op. cit.*, p. 53.
[3] *Op. cit.* pp. 58-59.
[4] In this discussion a price index is used for purposes of illustration. The treatment follows very closely that of Wesley C. Mitchell in *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 284.

FIGURE 86

Distribution of the Price Variations of 241 Commodities in 1913

(Percentages of Rise or Fall in Prices)



rising prices. This, of course, does not mean that all prices change in the same direction at the same time, nor that those which change together vary in the same degree.[1] All that is meant is that in terms of a single year or an average of years taken as a base, the price level moves up or down through relatively long periods. The differences of price from the norm, whether negative or positive, generally tend to be in the same direction. Large differences, of course, are less common than small ones, but those that are positive do not exactly compensate for those that are negative. Mitchell has shown this in a striking way by comparing the price variations of 241 commodities in 1913, computed, first, as percentages of rise or fall from the prices in 1912; and second, as percentages of rise or fall from the average prices of 1890-1899. Graphically, Figure 86 [2] shows the percentage changes of rise and fall.

The percentage differences—excesses and deficiencies of the 1913 prices relative to the 1912 prices—arrange themselves, as shown by the solid line, about a norm, the arithmetic mean, the mode and the median tending closely to agree.

"But the distribution of the second set of variations (percentages of change from the average prices of 1890-1899) as represented by the area inclosed within the dotted line has no obvious central tendency; it shows no high degree of concentration around the arithmetic mean (+30.4 per cent) or median (+26 per cent) and it has a range between the greatest fall (52.2 per cent) and greatest rise (234.5 per cent) so extreme that two of the cases could not be represented on the chart.

"Price variations, then, become dispersed over a wider range and less concentrated about their mean as the time covered by the variations increases. The cause is simple: With some commodities the trend of successive price changes continues distinctly upward for years at a time; with other commodities there is a consistent

[1] See Fisher, Irving, *op. cit.*, Chapter II for a discussion and for various graphic illustrations of the dispersion of the prices of 36 commodities, 1913 to 1918. See also Figure 65, *supra*, p. 335, showing price dispersion from 1891-1918.

[2] *Op. cit.*, p. 20.

TABLE 87

DISTRIBUTION OF 5578 CASES OF CHANGE IN THE WHOLESALE PRICES OF COMMODITIES FROM ONE YEAR TO THE NEXT, ACCORDING TO THE MAGNITUDE AND DIRECTION OF THE CHANGES

(Based upon the chain relatives in Table 11 of Bulletin of the Bureau of Labor Statistics, No. 149)

| Rising Prices | | | | | | Falling Prices | | |
|---|---|---|---|---|---|---|---|---|
| Per Cent of Change from the Average Price of the Preceding Year | Number of Cases | Proportion of Cases | Per Cent of Change from the Average Price of the Preceding Year | Number of Cases | Proportion of Cases | Per Cent of Change from the Average Price of the Preceding Year | Number of Cases | Proportion of Cases |
| 102–103.9 | 1 | 0.018 | 46–47.9 | 11 | 0.197 | Under 2 | * 405 | 7.261 |
| 100–101.9 | 1 | .018 | 44–45.9 | 10 | .179 | 2– 3.9 | * 375 | 6.723 |
| 98– 99.9 | — | — | 42–43.9 | 6 | .108 | 4– 5.9 | 329 | 5.898 |
| 96– 97.9 | — | — | 40–41.9 | 14 | .251 | 6– 7.9 | * 238 | 4.267 |
| 94– 95.9 | — | — | 38–39.9 | 17 | .305 | 8– 9.9 | 200 | 3.585 |
| 92– 93.9 | — | — | 36–37.9 | 11 | .197 | 10–11.9 | 173 | 3.101 |
| 90– 91.9 | — | — | 34–35.9 | 18 | .323 | 12–13.9 | * 120 | 2.151 |
| 88– 89.9 | — | — | 32–33.9 | 17 | .305 | 14–15.9 | 107 | 1.918 |
| 86– 87.9 | 1 | .018 | 30–31.9 | 22 | .394 | 16–17.9 | 76 | 1.362 |
| 84– 85.9 | 1 | .018 | 28–29.9 | 30 | .538 | 18–19.9 | 71 | 1.273 |
| 82– 83.9 | 1 | .018 | 26–27.9 | 29 | .520 | 20–21.9 | 45 | .807 |
| 80– 81.9 | 1 | .018 | 24–25.9 | 47 | .843 | 22–23.9 | 39 | .699 |
| 78– 79.9 | — | — | 22–23.9 | 45 | .807 | 24–25.9 | 32 | .574 |
| 76– 77.9 | — | — | 20–21.9 | 65 | 1.165 | 26–27.9 | 17 | .305 |
| 74– 75.9 | 1 | .018 | 18–19.9 | 73 | 1.308 | 28–29.9 | 27 | .484 |
| 72– 73.9 | 4 | .072 | 16–17.9 | * 102 | 1.828 | 30–31.9 | 16 | .287 |
| 70– 71.9 | 1 | .018 | 14–15.9 | 106 | 1.900 | 32–33.9 | 7 | .125 |
| 68– 69.9 | 3 | .054 | 12–13.9 | 115 | 2.062 | 34–35.9 | 10 | .179 |
| 66– 67.9 | 4 | .072 | 10–11.9 | 167 | 2.994 | 36–37.9 | 7 | .125 |
| 64– 65.9 | — | — | 8– 9.9 | * 237 | 4.249 | 38–39.9 | 5 | .090 |
| 62– 63.9 | — | — | 6– 7.9 | 261 | 4.679 | 40–41.9 | 5 | .090 |
| 60– 61.9 | 4 | .072 | 4– 5.9 | * 356 | 6.382 | 42–43.9 | 4 | .072 |
| 58– 59.9 | 6 | .108 | 2– 3.9 | 355 | 6.364 | 44–45.9 | 2 | .036 |
| 56– 57.9 | 1 | .018 | Under 2 | * 410 | 7.350 | 46–47.9 | 1 | .018 |
| 54– 55.9 | 3 | .054 | — | — | — | 48–49.9 | 1 | .018 |
| 52– 53.9 | 4 | .072 | No change* | * 697 | 12.494 | 50–51.9 | 1 | .018 |
| 50– 51.9 | 1 | .018 | — | — | — | 52–53.9 | — | — |
| 48– 49.9 | 5 | .090 | — | — | — | 54–55.9 | 1 | .018 |

SUMMARY

| | Number of Cases | Proportion of Cases |
|---|---|---|
| *Rising prices* | 2,567 | 46.021 |
| No change | 697 | 12.494 |
| Falling prices | 2,314 | 41.485 |
| Total | 5,578 | 100.000 † |

* Location of the deciles.
† *Op. cit.*, p. 18.

downward trend; with still others no definite long-period trend appears. In any large collection of price quotations covering many years each of these types, in moderate and extreme form, and all sorts of crossings among them, are likely to occur. As the years pass by the commodities that have a consistent trend gradually climb far above or subside far below their earlier levels, while the other commodities are scattered between these extremes. Thus the percentages of variation for any given year gradually get strung out in a long, thin, and irregular line, without a marked degree of concentration about any single point."[1]

*The tendency for price changes, calculated from year to year, to arrange themselves around a central position*—to conform to the "normal law of error"—has been worked out by Mitchell for the years 1891-1913 for 5578 cases. The price of each of more than 230 commodities during this period was expressed each year as a percentage of its price in the preceding year. The changes were then arranged in ascending order from the greatest decrease up through no change to the greatest increase. For the whole distribution deciles were then worked out for each year. With the changes arranged in this manner it is easy to measure the concentration about a norm, and to indicate the differences by successive deciles. Mitchell's table showing the dispersion, and his comments concerning it are given in the footnote on page 330.

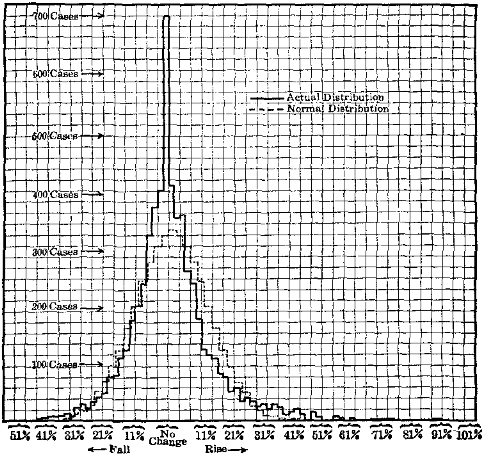
The actual distribution of the changes for the 5578 cases is given in Table 87, and is compared with a "normal curve of error" in Figure 87.

[1] *Op. cit.*, 21-22.

FIGURE 87

DISTRIBUTION OF 5578 PRICE VARIATIONS

(Percentages of Rise or Fall from Prices of Preceding Year)




In commenting upon the form of this distribution and its relation to the normal error curve, Mitchell says:

"There are several points to notice here. While the actual and the 'normal' distributions look much alike, they are not, strictly speaking, of the same type. The actual distribution is much more pointed than the other, and has a much higher 'mode,' or point of greatest density. On the other hand, the actual distribution drops away rapidly on either side of this mode, so that the curve representing it falls below the curve representing the 'normal' distribution. The actual distribution is 'skewed' instead of being perfectly symmetrical. The outlying cases of a 'normal' distribution extend precisely the same distance from the central tendency in both direc-

tions, whereas in the actual distribution the outlying cases run about twice as far to the right (in the direction of a rise of prices) as to the left (in the direction of a fall). This fact suggests that the actual distribution would be more symmetrical if it were plotted on a logarithmic scale, one which represents the doubling of one price by the same distance from zero as the halving of another price. Another aspect of the difference in symmetry is that the central tendency about which the variations group themselves is free from ambiguity in one case but not in the other. In the 'normal' distribution this tendency may be expressed indifferently by the median, the arithmetic mean, or the mode; for these three averages coincide. In the actual distribution, on the contrary, these averages differ slightly; the median and the 'crude' mode stand at $\pm 0$, while the arithmetic mean is $+1.36$ per cent. These departures of the actual distribution from perfect symmetry possess significance; but the fact remains that year-to-year price fluctuations are highly concentrated about their central tendency."[1]

The agreement between the distributions of price variations measured from year to year and the normal curve of error is important in the interpretation and calculation of index numbers. Many index numbers are of the average-of-relatives type. That is, relatives or ratios based upon a fixed or changing base are averaged in order to compare price changes from year to year. For this purpose the arithmetic mean is customarily used. But it is markedly affected by extremes. Accordingly, if the deviations from an average are not symmetrically distributed about a norm or central position, the arithmetic mean is a poor measure of central tendency. If, on the other hand, distribution is normal, or approximately so, as in the case of the chain-relatives shown in Figure 87, then the arithmetic mean agrees with or is not markedly different from the median and the mode, and may be used to describe, as accurately as any single amount can—*with this form of an index number*—the nature of price change.

Mitchell, after expressing price changes (1) on a remote fixed base—1890-1899—and (2) on a year-to-year base, concludes as follows:

[1] *Op. cit.*, pp. 18-19.

"The consequence is that the measurement of price fluctuations becomes difficult in proportion to the length of time during which the variations to be measured have continued. In other words, the farther apart are the dates for which prices are compared, the wider is the margin of error to which index numbers are subject, the greater the discrepancies likely to appear between index numbers made by different investigators, the wider the divergencies between the averages and the individual variations from which they are computed, and the larger the body of data required to give confidence in the representative value of the results."[1]

Two important questions are raised by the above discussion: (1) should reliance be placed in an average of relatives index number, and (2) if a relative is used, what average should be employed? These questions are discussed immediately below.

## V. The Methods of Constructing Index Numbers

Illustrations of three major methods of constructing index numbers are given above by using wholesale prices of paper in Chicago. Each of them needs to be considered separately.

### 1. Averages of Relatives (Ratios)

#### (1) *Fixed vs. Shifting Base*

In order to compute an average of relatives, a base must be selected in terms of which to express the prices as percentages. In making a choice, two alternatives are presented, (1) a single year which is made common to all the series,[2] and (2) the preceding year changing from year to year. The first is known as a fixed base; the second as a shifting base. When relatives are computed in terms of a fixed year, the index is known as a "fixed base relative"; when in terms of a shifting year, and the resulting ratios are multiplied together, the number is known as a "chain-relative."

Table 79 shows such a fixed base relative—in terms of 1913

[1] *Op. cit.*, p. 22.

[2] In some cases, the average price during a series of years is used. The base, however, is fixed—that is, it applies to all of the years.

—the arithmetic mean, the median, and the geometric mean being the averages in which the various changes are measured. Table 80 shows a chain-relative calculated upon the same base period.

### a. Arithmetic Means of Relatives—Fixed Base

In computing a fixed base relative, some year or number of years which is thought of as normal is selected. By computing the prices as percentages of this base, differences in prices as well as in the units in which the prices are quoted are supposed to be reduced to a common denominator so that they can be totaled and averaged. But as has been shown, the dispersion of relatives computed upon a fixed base, more particularly when it is remote, is large, and the distribution skewed in the direction in which most prices are moving. Arithmetic averages of relatives do not, under these conditions, reflect the typical or modal movement. They are too much affected by the extremes.

Moreover, the importance or weight assigned to the amount of the change is inversely proportional to the magnitude of the price in the base year. If prices change, dividing them by the base price does not bring them to a comparable basis, unless they all change at the same rate—which they do not do in the case of the wholesale prices of paper, nor with the prices used by Mitchell. Indeed, it is safe to say that uniformity of change is never encountered. To add and take an arithmetic mean of relatives gives too much weight to increasing and too little weight to decreasing prices. Moreover, more weight is given to rapidly rising than to slowly rising prices, and more to rapidly falling than to slowly falling prices.

### b. Medians of Relatives—Fixed Base

But medians of fixed base relatives may be used rather than arithmetic means. What may be said in their favor? Medians are less affected by extreme items than are arithmetic

means, and, therefore, are likely to be more typical of price changes. But (1) there may be no *actual* median items;[1] (2) medians of different groups cannot be combined nor averaged;[2] (3) they are not reversible, that is, index numbers based upon them cannot be shifted from base to base by division;[3] and (4) they are erratic when there are few items.[4] Moreover, to take medians of relatives does not remove the bias to which relatives are due in periods of rising and falling prices. The bias here is due to the method of measuring the change, not to the method of averaging it.

### c. Geometric Means of Relatives—Fixed Base

Instead of using arithmetic means or medians of relatives, geometric means may be employed. If the *average ratio of change in prices is to be measured, the geometric mean should* be used. This average gives equal influence to equal ratios of change, irrespective of the previous level of the prices, production, stocks, or what not to which the changes apply. The doubling of one price, for instance, is exactly counterbalanced by the halving of another when a geometric average of the changes is taken. Accordingly, geometric means are always smaller than arithmetic means of relatives.[5] They may be smaller or greater than medians of relatives.

An illustration will help to make the distinction clear between measuring price changes by an arithmetic mean of relatives and by a geometric mean of relatives.

[1] Six of the medians had to be interpolated for in Table 79. While few items are involved in this illustration, the difficulty encountered is typical of medians. It does not occur *only* when few items are used. See the discussion of the median and of interpolation, pp. 286-289.

[2] This follows because, in order to locate medians, items must be arranged in order of magnitude.

[3] This follows because with a new base the *order* of the items will probably be different, therefore, giving a new median.

[4] See the medians in Table 79 which are located by interpolation.

[5] This condition would obtain in the illustration in Table 79 if full account were taken of decimal amounts.

| | Actual Prices | | Relative Prices | |
|---|---|---|---|---|
| *Commodity* | *First Year* | *Second Year* | *First Year* | *Second Year* |
| A | $1.00 | $2.00 | 100 | 200 |
| B | .50 | .25 | 100 | 50 |

Change measured by

| (1) The Arithmetic Mean of Relatives | *First Year* | *Second Year* | (2) the Geometric Mean of Relatives. *First Year* | *Second Year* |
|---|---|---|---|---|
| Sum of Relatives = | 200 | 250 | | |
| Average of Relatives = | 2)200 100 | 2)250 125 | $\sqrt{100 \times 100} = 100$ | $\sqrt{200 \times 50} = 100$ |
| Index Number = | 100 | 125 | Index Number = 100 | 100 |

Measured by the arithmetic mean of relatives, prices rose 25 per cent; by the geometric mean, they remained the same.

Moreover, as pointed out by Mitchell, the geometric mean "is not in danger of distortion from the asymmetrical distribution of price variations."[1] This fact is of real significance since distributions of price fluctuations are skewed either positively or negatively—positively during periods of rising prices, and negatively during periods of falling prices—when calculated on any other than a year-to-year base.[2] Accordingly, geometric means are closer to the modal change than are arithmetic means, and the modal or typical change is of primary interest when speaking of *the* change in prices.

### *(2) Chain-Relatives*

The distribution of relative prices calculated on the preceding year as a base conforms more closely to the normal curve of error than does that made from relatives computed on

[1] *Op. cit.* p. 69.

[2] See *op. cit.*, p. 70, for a table showing the positive skewness of the relative prices of 1437 commodities in 1918 on the base—July, 1913, to June, 1914.

a remote fixed base. If the relative or percentage method is to be used to measure price change, then a near base is to be preferred to one that is distant. Accordingly, link-relatives, which are later placed in a chain, are sometimes used for this purpose. But it is not easy to give a precise meaning to such a chain except at adjoining links.

When, for instance, the index number for paper prices in Chicago in 1921 is linked up through all of the changes from 1913 to 1921, one is in doubt as to exactly what it measures. This method, however, make it possible to drop old and to add new commodities—a necessity frequently encountered when computing a series of numbers over a period of years. But as Mitchell shows, full agreement in price change is not to be expected by the use of the fixed and the chain base methods.[1]

### (3) *Base Shifting and the Use of Averages of Relatives*

#### a. When Arithmetic Averages of Relatives are Used

In order to shift the base when arithmetic means of relatives are used, two methods are available: (1) recomputing the relatives of each commodity on the new base and averaging their sum—that is, reconstructing the number; and (2) shifting by the "short-cut" method. The first method gives a number having all the properties of the old one but expressed in another year as unity. The second method—which consists in dividing the index number for other dates by the figure chosen as the base—produces results which will not necessarily agree with those which would be secured if relatives were computed for each commodity on the new base. As Mitchell says,

"... For such recomputation usually alters considerably the relative influence exercised upon the arithmetic means by the price

[1] Mitchell, W. C., *Bulletin* 284, pp. 87-89. Compare the results in Table 86 secured by the fixed-base-relative and the chain-relative methods.

fluctuations of certain commodities. Those articles which are cheaper in the new than in the old base period get higher relative prices and, therefore, increased influence. Vice versa, articles that are dearer in the new base period get lower relative prices and, therefore, diminished influence. Of course the short method of shifting the base, which retains the old relative prices, does not permit any such alteration in the influence exercised by the fluctuations of different commodities. Hence the two methods of shifting the base seldom yield precisely the same results. To present a series of arithmetic means shifted by the short method as showing what the index numbers would have been if they had been computed upon the new base is, therefore, misleading."[1]

### b. When Medians of Relatives Are Used

When medians of relatives are used, shifting to a new base is impossible without recomputing the relatives for the individual commodities.[2]

### c. When Geometric Means of Relatives Are Used

Index numbers based upon geometric means of relatives can be shifted from base to base without error. The same result is secured by recomputing the commodity relatives and by dividing by the new index base figure. An illustration will make this clear.

Suppose the prices of two commodities were as follows:

| | Actual Prices | | | Relative Prices (1923 = 100) | |
|---|---|---|---|---|---|
| *Commodity* | *1923* | *1924* | | *1923* | *1924* |
| A | $1.00 | $2.00 | | 100 | 200 |
| B | 1.00 | .50 | | 100 | 50 |
| | | | Geometric Means = | $\sqrt{100 \times 100} = 100$ | $\sqrt{200 \times 50} = 100$ |
| | | | Index Numbers = | 100 | 100 |

[1] Mitchell, W. C., "Index Numbers of Wholesale Prices in the United States and Foreign Countries," *Bulletin* 173, *United States Bureau of Labor Statistics*, July, 1915, p. 39. See also the revision of this bulletin, Number 284, pp. 83-85.

[2] See the discussion of *Medians of Relatives—Fixed Base*, pp. 497-498.

Changing the base to 1924:

| (1) by recomputing the relatives and | | | (2) by dividing by the new base figure |
|---|---|---|---|
| | *1923* | *1924* | |
| | 50 | 100 | |
| | 200 | 100 | $\frac{1923}{1924} = \frac{100}{100} \times 100 = 100$ |
| | $\sqrt{50 \times 200} = 100$ | $\sqrt{100 \times 100} = 100$ | |
| Index Numbers: | 100 | 100 | Index Numbers: 1923 = 100; 1924 = 100 |

2. RATIOS OF AVERAGE PRICES

### *(1) Merits of the Method*

The ratios of arithmetic averages of actual prices—the units in which the quantities are priced being the same—do not have the bias inherent in arithmetic averages of relative prices, yet they are affected by the fact that the price for the same unit varies widely from commodity to commodity. For instance, the price in 1913 of "fine" paper is more than three times as important in determining the average (or the total) price for that year as is the price of "newspaper" for the same quantity. If the same proportions from year to year obtained among the different prices, the bias from this source would not enter. But they do not as is evident from an inspection of Table 79. An "unweighted" ratio of averages index number accordingly is arbitrarily weighted.[1]

If the unit in which the prices were taken varied, then another occasion for bias would enter, because the price would in part depend upon the unit. For instance, if "newspaper" were quoted in tons, the price would be increased enormously and the averages for the different types be largely controlled by it. At least one of the early index numbers was made by totaling the prices of articles quoted in their customary commercial units.

[1] See the discussion of Bradstreet's Index Number, Chapter XVI, pp. 523-525.

### (2) *Methods of Base Shifting Illustrated*

Inasmuch as actual prices are averaged or totaled—the price quotations having been reduced to the same unit—no base period is involved. Any one of the years, however, may be chosen as a base and the average or total price for each of the other years be expressed as a percentage of it. Moreover, the base can be shifted from year to year without error, provided the prices refer to the same source through the period. An illustration will show that this is the case.

Price of Newsprint per 100 lbs.

| Jobber | 1913 | 1920 | 1921 |
|---|---|---|---|
| A | $3.00 | $11.00 | $8.00 |
| B | 3.25 | 11.20 | 8.40 |
| C | 3.50 | 10.90 | 8.70 |
| D | 2.75 | 12.10 | 8.30 |
| Total Price | $12.50 | $45.20 | $25.40 |
| Average Price | 3.125 | 11.30 | 6.35 |
| Relatives—(1913 = 100) | 100 | 361.6 | 203.2 |

It is desired to shift the base from 1913 to 1921. This may be done (1) by expressing the prices in 1913 and in 1920 as percentages of the price in 1921. The results by this method are as follows:

| 1913 | 1920 | 1921 |
|---|---|---|
| 49.2 | 178.0 | 100.0 |

or (2) by multiplying through, thus

$$1920 \text{ on } 1921 \text{ base, } \frac{361.6}{203.2} \times 100 = 178.00$$

$$1913 \text{ on } 1920 \text{ base, } \frac{100.0}{361.6} \times 100 = 27.65$$

Therefore, 1913 on 1921 = 178.0 × 27.65 = 49.2, which is the same result as is secured by the first method.

## 3. WEIGHTED AGGREGATES OF ACTUAL PRICES AND BASE SHIFTING

### (1) *Method of Computation and Relative Merits*

The recent developments in the making of index numbers have been toward the use of aggregates of actual prices weighted by suitable quantities. The method consists in (1) applying to the *price* of each commodity a *quantity* weight indicative of its importance, (2) totaling the products, and (3) expressing the results in the form of relatives on a base period. It was by this method that the index numbers for wholesale paper prices in Table 85 were computed.[1]

The advantages claimed for index numbers computed by this method may be summarized as follows: (1) they are easy to understand; (2) easy to compute; (3) do not require a base period for the calculation of relatives, but may be placed on a relative basis after the products are computed and totaled; (4) the base can be shifted at will without error; (5) they are not distorted during periods of rapid price change; and (6) they measure the change in the money cost of goods —the end most frequently desired from the use of an index number.

### (2) *Methods of Base Shifting Illustrated*

The claim that the base in weighted aggregates of actual prices can be shifted at will without error needs to be demonstrated. For this purpose the index numbers calculated for paper prices in Chicago (Table 85) may be used as an illustration. The index numbers in the table are based on 1913. It is desired to shift the base to 1921. This may be done by dividing each of the price aggregates by the amount for the new base year. To illustrate: The index for 1919 on 1921, $\frac{1037.9}{1155.5} \times 100$, is 90; that for 1913 on 1921, $\frac{508.3}{1155.5} \times 100$, is 44.

[1] See the discussion of the index numbers of the United States Bureau of Labor Statistics, Chapter XVI, pp. 516-518.

With 1913 as the base, the index for 1919, $\frac{1037.9}{508.3} \times 100$, is 204.

From the formulæ for 1919 on 1921, $\frac{1037.9}{1155.5} = 90$, and for 1919 on 1913, $\frac{1037.9}{508.3} = 204$, it is possible to get the index for 1913 on 1921 by simple division. Thus, $\frac{1037.9}{1155.5} \div \frac{1037.9}{508.3} = \frac{508.3}{1155.5} = \frac{90}{204}$. That is, $\frac{90}{204} \times 100 = 44$, which is the index of 1913 on the base of 1921.

## VI. Weighting

### 1. MEANING AND METHODS OF WEIGHTING

Distinction is generally made between weighted and unweighted index numbers, but often without a clear idea of what is meant by the terms. Every index number is weighted in some form. So-called "unweighted" series are generally haphazardly weighted; while in those which are termed weighted, the weights are selected according to some systematic plan.

If the average of relatives method is used, each item being counted once, the explicit[1] weights are unity in each case. If, on the other hand, the weighted average of relatives method is followed, the weights are the *values* applied to each relative in order to secure the products to be averaged. On the other hand, if the weighted aggregate of actual prices method is used, the weights are the *quantities* which are applied to the actual prices in order to get the products which are totaled into aggregates and later placed on a relative base.[2]

Weighting is effected in either of two ways: the first method

[1] Defined below.

[2] To weight prices by values is illogical because the values in this case are the results of multiplying quantities by prices.

is in the selection of the commodities themselves—varying emphasis being given to the different items by the number of times a given article or one of the same general class is included. This may be called the "implicit" method. The second way is to use some outward evidences of importance—that is, to apply "explicit" weights.

The explicit weights commonly assigned to retail prices in computing an index designed to measure changes in the cost *of living, are the quantities of the articles consumed.* Similarly, the weights applied to wholesale prices, in the construction of an index to show general changes in prices, are the *total amounts of goods placed on the market,* aggregate expenditures by the people of a country, values produced, values consumed, values exchanged computed at the price in the year the level of which is in question, etc. If the changes in prices which are being considered apply to securities rather than to commodities, then suitable explicit weights for different purposes might be the amounts outstanding, the earnings of the companies to which the securities apply, the dividend rates, etc. But the use of these different systems of weights produces different results. So we are brought back to the question: What is it that weights are intended to do?

Lack of attention to weights does not mean that weights are equal, but generally that they are haphazard. They are not necessarily bad because of this, nor good, as Mitchell points out, if they are consciously made. "The real problem for the maker of index numbers is whether he shall leave weighting to chance or seek to rationalize it."[1]

Moreover, so-called unweighted index numbers may in fact be markedly weighted by the use of "implicit" weights; as, for instance, in the Aldrich index number, where 25 different varieties of pocket knives were included, thus "giving this trifling article an influence upon the result more than eight times greater than given to wheat, corn, and coal put together."

[1] *Op. cit.*, p. 60.

Truly to give each commodity equal weight requires careful and studied attention to the choosing of positive weights.

But what test or tests of importance are available? Are they applicable at all times and places, and for all purposes? To weight a retail price index number—where the purpose of its computation is to measure the effect of price change on consumers—by the amount of production or by the value of the articles exchanged is ill fitting. Likewise, to weight wholesale prices by statistics of family consumption is illogical. Weights should be appropriate or they should be dispensed with entirely.

On the relation of weights to purposes of index numbers, Mitchell says:

"If rational weighting is worth striving after, then by what method shall the weights of the different commodities be arrived at? That depends upon the object of the investigation. If, for example, the aim be to measure changes in the cost of living, and the data be retail quotations of consumers' commodities, then the proportionate expenditures upon the different articles as represented by collections of family budgets make appropriate weights. If the aim be to study changes in the money incomes of farmers, then the data should be 'farm prices,' the list of commodities should be limited to farm products, and the weights should be proportionate to the total money receipts from the several products. If the aim be to construct a 'business barometer,' the data should be prices from the most representative wholesale markets, the list should be confined to commodities whose prices are most sensitive to changes in business prospects and least liable to change from other causes, and the weights may logically be adjusted to the relative faithfulness with which the quotations included reflect business conditions. If the aim be merely to find the differences of price fluctuation characteristic of dissimilar groups of commodities, or to study the influence of gold production or the issue of irredeemable paper money upon the way in which prices change, it may be appropriate to strike a simple arithmetic average of relative prices. If, on the other hand, the aim be to make a general-purpose index number of wholesale prices, the question is less easy to answer."[1]

[1] *Op. cit.*, pp. 62-63.

But why use weights at all, when weighted results are so strikingly the same as unweighted? Two main reasons are usually assigned for ignoring them: (1) the difficulty of finding suitable weights and of currently correcting them, and (2) the fact that unweighted series are almost identical with those which are weighted. Bowley, in much quoted passages, says:

"The discussion of the proper weight to be used . . . has occupied a space in statistical literature out of all proportion to its significance, for it may be said at once that no great importance need be attached to the special choice of weights; one of the most convenient facts of statistical theory is that, given certain conditions, the same result is obtained whatever logical system of weights is applied."[1]

"So we arrive at a very important precept; *in calculating averages give all care to making the items free from bias, and do not strain after exactness in weighting.*"[2]

But this is hardly a full statement of the case. Properly to weight a number is to make it "free from bias." This may be done by assigning weights to the samples at hand or by the more direct, but sometimes more difficult, method of choosing more samples. In reality the two are alternatives, with this difference that errors in prices will probably tend more nearly to be compensating than those in weights. If a rational system of weights does not change the result of an "unweighted" average, then weights may be dispensed with; if it does, then they ought to be used.

While the problem of selecting weights lends itself to theoretical discussion, it is primarily of practical concern. To the person who desires to use index numbers the question cannot be dismissed with the assertion that if weights are chosen according to chance, weighted and unweighted indexes closely agree. As they are computed, weights are not always so chosen, numbers differ materially, and the merits of unweighted and weighted numbers can be determined only by

[1] Bowley, A. L., *Elements of Statistics*, 2d Ed., 1902, p. 113.
[2] *Ibid.*, p. 118.

comparison.[1] In the light of the differences shown in this manner the merits of the two types of series must be determined. The student and the business man cannot readily make these comparisons for themselves but they can be familiar with those that have been made. That "amiable weakness to take upon faith plausible figures that fill a pressing want" would not then be so common.

Should weights be fixed or fluctuating? By changing them a more accurate measure of importance is undoubtedly acquired, but changes in an index must then be interpreted not only in terms of prices but also in terms of weights. Conceivably, some sort of an average of relative importance over a period could be used, but if so, the variations would be lost sight of. When chain-indexes are used, weights can be varied without confusion, since price changes from year to year only are measured. Such figures do not accurately measure changes over a period.

### 2. WEIGHTING IN PROFESSOR FISHER'S "IDEAL" FORMULA

Professor Fisher[2] by an elaborate analysis of the types of bias by which index numbers computed from averages of relatives of different kinds, and from aggregates of actual prices are affected, concludes that a scheme of cross weighting should be used. In this manner he claims to overcome the types of bias by which prices and quantities are affected. He writes his formula as follows:

$$\sqrt{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}\right) \times \left(\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right)}$$

[1] Weighted and unweighted series, and those weighted in various ways both for commodities and stocks, are elaborately compared by Mitchell, Wesley C., in "Critique of Index Numbers of Prices of Stocks" in *The Journal of Political Economy*, July, 1916, *passim;* and *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 173, pp. 74-75. See also Fisher, Irving, *op. cit.*, where the effects of applying weights are worked out in great detail.

[2] Fisher, Irving, *op. cit.*, *passim.*

where $\Sigma$ = "the sum of such terms as"

$p_1$ = the price of any commodity in a given year or other period.

$q_1$ = the quantity of the commodity in the given years or other period.

$p_0$ = the price of any commodity in the base year or other period.

$q_0$ = the quantity of that commodity in the base year or other period.

This formula requires both price and quantity (weights) for each year to which an index applies. As will be noted, there are four sets of aggregates required: (1) prices in the given year multiplied by quantities in the base year; (2) prices in the given year times quantities in the given year; (3) prices in the base year times quantities in the base year; and (4) prices in the base year times quantities in the given year. The first and second aggregates are divided by the third and fourth aggregates, respectively, giving two relatives which are then multiplied together, and the square root of the product extracted.

In this formula—Fisher calls it the "Ideal" because it most fully neutralizes the types of bias which he finds in measuring changes in prices and in quantities—the form of weighting is designed so that the index number secured will meet two basic tests: viz., "time reversal" and "factor reversal." The time reversal test Fisher describes as follows:

"The test is that the formula for calculating an index number should be such that it will give the same ratio between one point of comparison and the other, *no matter which of the two is taken as the base.*

"Or, putting it another way, the index number reckoned forward should be the reciprocal of that reckoned backward."[1]

By this he means that if an index shows that between 1913 and 1920, for instance, prices doubled, then it should show

[1] *Op. cit.*, p. 64.

that the level in 1913 was one-half of that in 1920 when measured from the latter year.

Concerning the "factor reversal" test he says:

> "Just as our formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results—*i.e.*, the two results multiplied together should give the true value ratio."[1]

It is unnecessary to enter into a discussion of the merits of this particular formula, or the question as to whether there is one formula which is best—"ideal"—for all purposes.[2] It suffices for our purposes to call attention again to the fact that the peculiar cross weighting is advised largely because it equalizes different types of bias, thus definitely associating rather than contrasting "making the items free from bias" and "straining after exactness in weighting."

All index numbers are no longer considered to be equally good. Study of the methods of their construction, of the price fluctuations of different types of commodities, of bias, etc., has made the maker of index numbers critical. He is no longer satisfied with the crude methods of yesterday in the face of the specific findings of such students as Mitchell and Fisher. How about the attitude of the user? He is not so critical, but he should be. After all, it is he who applies the numbers to the different problems which he has to solve. It may be worth while, therefore, to offer in brief form some suggestions which will help him to make a discriminating application.

## VII. Suggestions to Users of Price Index Numbers

1. Before applying index numbers to specific problems, clearly formulate a statement of the use which you have in mind.

[1] *Op. cit.*, p. 72.

[2] Its form, ease of calculation, suitability to different purposes, etc., have been the subject of vigorous controversy. See, for instance, Fisher, *op. cit.*, *passim*; Persons, W. M., *Review of Economic Statistics*, Preliminary, Volume 3, pp. 103-113 (May, 1921); Mitchell, *op. cit.*, pp. 91-93, and the references given.

In doing this it will be necessary to distinguish, among other things, between

(1) a general and a specific use.
(2) different general uses.
(3) different specific uses.

2. Distinguish between index numbers designed to measure changes in the prices of
   (1) commodities sold at wholesale and at retail.
   (2) manufactured products and raw materials.
   (3) basic commodities in central markets, and farm products, for instance.
   (4) foods and the "total cost of living."
   (5) goods with business "barometric" significance and those relating, for instance, to consumption.
3. Observe the methods according to which index numbers are constructed, paying special attention to
   (1) the kinds of commodities included.
   (2) the source of information on prices.
   (3) the nature of the prices—as market, contract, import and export.
   (4) the number of commodities.
   (5) the kinds of weights used, and the source of information.
   (6) the periods to which the index applies.
   (7) the base period, if any.
   (8) the type of average used, as arithmetic mean, median, geometric mean.
4. Avoid
   (1) shifting the base by the "short-cut" method when arithmetic means and medians of relatives are used.
   (2) confusing long and short period price trends.
   (3) confusing numbers which measure average ratios of change in price, and average change in amount of money required to buy a bill of goods.

(4) confusing an index number of price and its reciprocal, the purchasing power of the dollar.

5. Choose the index number which most fully meets the needs in your particular case, but do not use it blindly. *An index number shows what it shows* and nothing else. What this is should and can be known by the user.

## VIII. Conclusion

In this chapter our aim has been (1) to show by concrete examples the different methods of constructing index numbers, (2) to explain and briefly to criticize each of the methods, and (3) to offer some helpful suggestions to users of index numbers. Little more, however, has been done than to touch upon the more important phases of the subject. Students should consult the painstaking studies of Fisher, Mitchell, and others if they wish really to understand the subject.

This chapter is not a critique, but rather an exposition of the principles upon which a critique must be based. If an interest in index number making and using has been aroused, the main purpose of what has been written here will have been accomplished. After all, chief reliance must be placed in the scientific spirit and integrity of both maker and user. If these are lacking, the use of statistics is without a logical defense.

### References

Barnett, G. E., "Index Numbers of the Total Cost of Living," *Quarterly Journal of Economics*, February, 1921, pp. 240-263.

Bowley, A. L., *Elements of Statistics*, 4th Edition, King, London, 1920, Chapter IX, pp. 196-214.

Davies, G. R., "The Problem of a Standard Index Number Formula," *Journal of the American Statistical Association*, June, 1924, pp. 180-188.

Fisher, Irving, *The Purchasing Power of Money*, Macmillan, New York, 1911, Chapter 10.

Fisher, Irving, *The Making of Index Numbers*, Houghton Mifflin, Boston, 1922, especially Chapters II, III, IV, X, and XI.

FLUX, A. W., "The Measurement of Price Changes," *Journal of the Royal Statistical Society*, March, 1921.

HOOKER, R. H., "The Course of Prices At Home and Abroad, 1890-1910," in *The Journal of the Royal Statistical Society*, Vol. LXXV, pp. 1-36 (December, 1911).

MACAULAY, F. R., "The Making and Using of Index Numbers," *American Economic Review*, December, 1915, pp. 928-931.

MEEKER, ROYAL, "Some Features of the Statistical Work of the Bureau of Labor Statistics," in *The Quarterly Publications of the American Statistical Association*, March 1915, pp. 431-441.

MILLS, FREDERICK C., *Statistical Methods Applied to Economics and Business*, Holt, New York, 1924, Chapter VI, pp. 169-251.

MITCHELL, W. C., *Business Cycles*, University of California Studies, Berkeley, 1913, pp. 112-139, on "The Representative Character of Index Numbers."

MITCHELL, W. C., "Index Numbers of Wholesale Prices in the United States and Foreign Countries," United States Department of Labor, *Bulletin Bureau of Labor Statistics*, Whole No. 284, Oct., 1921. Part I, "The Making and Using of Index Numbers," pp. 5-114; Part II, "Index Numbers of Wholesale Prices in the United States and Foreign Countries," pp. 115-343. (*This publication in the field of index numbers is epoch-making.* It includes a complete study of the technique of the construction and use, as well as a descriptive account of current and past index numbers in this and in foreign countries.)

PERSONS, W. M., "Fisher's Formula for Index Numbers," *The Review of Economic Statistics*, Preliminary Volume 3, No. 5, May, 1921, pp. 103-113.

SNIDER, JOSEPH L., "Wholesale Prices in the United States, 1866-1891," *The Review of Economic Statistics*, Volume VI, No. 2, April, 1924, pp. 93-118.

SNYDER, CARL, "A New Index of General Price Level," *Journal of the American Statistical Association*, June, 1924, pp. 189-195.

YOUNG, A. A., "Index Numbers," in Rietz, H. L. (Editor-in-Chief), *Handbook of Mathematical Statistics*, Houghton Mifflin, Boston, 1924, pp. 181-194.

YOUNG, A. A., "The Measurement of Changes of the General Price Level," *Quarterly Journal of Economics*, Vol. 35, 1921, pp. 557-573.

# CHAPTER XVI

# PRICE, QUANTITY, AND GENERAL BUSINESS INDEXES DESCRIBED AND COMPARED

## I. Introduction

The purpose of the preceding chapter was to illustrate the different methods by which index numbers of prices and of other phenomena may be computed, and to discuss the principles involved. The purpose of this one is to describe and compare the methods used in the more important public and private series.

The treatment is of necessity brief. It includes only an outline of the methods peculiar to each type. While the facts presented are for the most part readily available, they are not generally kept in mind when index numbers are used. It may be helpful to the reader, therefore, to have at hand a brief account of the more important series.

## II. Index Number of Prices

American commodity [1] price index numbers may be divided into two groups: (1) those prepared by agencies of the United States Government, and (2) those issued by private organizations. The more commonly known indexes from both sources are described in what follows:

[1] Excellent summaries under the headings, among others—history, source of quotations, base period, number and class of commodities, grouping, weighting, etc.—of foreign price index numbers are contained in *Bulletin* 284 *of the United States Bureau of Labor Statistics*, pp. 175-336.

## 1. PRICE INDEX NUMBERS ISSUED BY THE UNITED STATES GOVERNMENT

### (1) *Index Numbers of Wholesale Prices*

#### a. The United States Bureau of Labor Statistics' Wholesale Price Index Number[1]

The systematic publication of a wholesale price index number by the United States Government was begun in 1902. The period first covered was 1890 to 1901, inclusive. This number was in continuation of the index compiled by the Department of Labor for the period 1890 to 1899, but included somewhat different commodities and carried the computations back to 1890. Since then, monthly and annual numbers have appeared regularly.

Up to and including 1913, the index number was an average of relatives based upon the average price, 1890-1899. In 1914 a change was made to an aggregate of actual prices weighted according to the amount of goods placed on the market in 1909. The weights now used are the amounts of goods marketed in 1919.

The change from an average of relatives to a weighted aggregate of actual prices method was made primarily because of (1) the difficulty of changing the base in averages of relatives without entirely recomputing the series; (2) a realization that an arithmetic average of relatives does not accurately measure typical price changes, more especially during periods of rapidly rising prices;[2] and (3) the conviction that a price series built up from actual money prices shows most accurately what the Bureau wanted to show—changes in the cost of "an unvarying market basket."

The important details about the method now used by the

[1] For a complete description of this index, see *Bulletin of the United States Bureau of Labor Statistics* No. 326, Washington, D. C., March, 1923.

[2] See the discussion of *Dispersion of Price Fluctuations*, *supra*, p. 489 *ff.*

Bureau of Labor Statistics in computing its wholesale price index number are as follows:

(a) The Price Quotations

Prices of 450 commodities, obtained primarly from trade journals, manufacturers, sales agents, trade bodies, etc., are collected systematically and regularly by the Bureau. Contact with the trade, a carefully prepared system of record cards providing methods for establishing the identity of commodities, and editorial care guarantee substantial accuracy of the prices secured. So far as possible, the quotations are secured weekly from primary markets.

(b) Types and Grouping of Commodities

The 450 commodity quotations are divided into the following groups—the numbers in parentheses representing the proportions falling in each group: farm products (12.4); foods (23.3); cloths and clothing (15.6); fuel and lighting (4.4); metals and metal products (11.8); building materials (10.4); chemicals and drugs (9.6); house furnishings (6.9); miscellaneous [1] (5.6).

(c) The Method of Calculating the Index

The average price [2] of each article for each year—404 rather than 450 are used in the index—is multiplied by the estimated quantity of the article marketed in the census year 1919—the amount in each case being checked against all available information. The products for the different commodities obtained in this manner are then added together. These different computations give a series of values from which the index number for each year is calculated as a relative or percentage number, the value for 1913 being taken as a base or 100.

[1] Cattle feed, leather, paper and pulp, other miscellaneous.

[2] Average yearly prices are built up from average weekly and monthly prices.

(d) The Form and Place of Publication

Monthly and annual index numbers for the commodity groups separately and combined, and reduced to relatives on the base, 1913, appear in *The Monthly Labor Review,* and in *Wholesale Prices,* both issued by the Bureau of Labor Statistics, Washington, D. C.

b. The Federal Reserve Board's Wholesale Price Index Number[1]

An index number of wholesale prices has been prepared by the Federal Reserve Board since October, 1918—the series being computed back to 1913.

(a) The Price Quotations

The price quotations are the same as those used by the United States Bureau of Labor Statistics in its wholesale series.

(b) Types and Grouping of Commodities

The commodities used are the same as those which make up the wholesale index of the United States Bureau of Labor Statistics, but they are grouped into three major classes, as follows: (1) raw materials, this group being further divided into farm products, animal products, forest products, and mineral products; (2) producers' goods; and (3) consumers' goods.

(c) The Method of Calculating the Index

The method of calculation is the same as that used by the Bureau of Labor Statistics, that is, a weighted aggregate of actual prices, the weights being the estimated quantities of goods marketed in 1919.

[1] For a complete description of this index number see *Bulletin* 284 *of the United States Bureau of Labor Statistics*, Washington, D. C., October, 1921, pp. 133-135.

(d) The Form and Place of Publication

Monthly and annual index numbers by commodity groups reduced to relatives on the base, 1913, appear monthly in *The Federal Reserve Bulletin,* Federal Reserve Board, Washington, D. C.

c. The United States Department of Agriculture's Wholesale Price Index Number of Farm Prices of Crops and of Livestock[1]

(a) The Price Quotations

The prices of the 30 commodities used in this index are those paid to producers as reported to the *Division of Crop and Market Estimates* of the Department. The prices refer to the 15th of each month.

(b) The Types and Grouping of Commodities

The prices cover (1) grains, (2) fruits and vegetables, (3) meat animals, (4) dairy and poultry, (5) cotton and cottonseed, and (6) unclassified.

(c) The Method of Calculation

An average price for each commodity for the period August 1909 to July 1914 is determined. The price for each commodity is then multiplied by the average quantity of the corresponding commodity marketed in the period 1918 to 1923, and the resulting values added together to form an aggregate value for the base period. This is taken as 100. Similar aggregates are computed for each month and year, and expressed as relatives or percentages of the aggregates of the base period.

[1] A full description of this index is contained in *Crops and Markets, Monthly Supplement,* United States Department of Agriculture, August, 1924, p. 285.

(d) The Form and Place of Publication

This index number by months and years and by groups of commodities appears in the *Monthly Supplement, Crops and Markets* of the Department.

### (2) *Index Numbers of Retail Prices*

If the collection of price data as a basis for the computation of a wholesale price index presents difficulties, as it undoubtedly does, these are many times more serious in the case of price data for a retail price index. While retail prices may change more slowly than wholesale prices, may be less affected by trade disturbances, and may move further in either direction after they are disturbed and be slower to regain their former position, it is these conditions and others, which make it so difficult to procure satisfactory price data over a period of time so as to measure the changes actually taking place. Prices of some commodities fluctuate from day to day; others less susceptible to conditions of demand and supply show appreciable change within somewhat longer periods. Prices of the same commodity vary materially as between localities. Some commodities, standard in character, but peculiar to local markets and not possessing distinctive trade names, sell at widely different prices at the same time.

#### a. The United States Bureau of Labor Statistics' Index Number of Food Prices

(a) The Price Quotations

From 1890 to 1907, the Bureau used 30 commodities. From 1907 to 1913, this number was reduced to 15, and in 1914 and 1915, respectively, the number was 17 and 21. Forty-three products are now used.

Prices of these commodities, on the 15th of each month in most cases, are secured from retailers in 51 cities of the United States. The prices are taken as representative of food products generally.

(b) Types of Commodities

The 43 articles are distributed into 22 groups for the purpose of computing index numbers of price change.

(c) The Method of Calculating the Index

From the monthly quotations, the Bureau computes an average price for each article in each city, and in the 51 cities combined. From these, relative prices or index numbers are computed for each article on the 1913 base price. For the index numbers showing prices in a city and for the United States as a whole, the prices are weighted according to the quantity of each article consumed by an average family during one year. The consumption weights (quantities) were secured from a comprehensive study made by the Bureau in 1918-1919.

(d) The Form and Place of Publication

Index numbers showing changes in food prices for groups of commodities, and for all articles combined, for the country as a whole appear in the *Monthly Labor Review*. From time to time, they are also shown separately by cities.

b. The United States Bureau of Labor Statistics' Index Number of Cost of Living

An index number showing the "changes in the cost of living" has been published by the Bureau of Labor Statistics since 1918, although the data go back to December, 1914. This index is a composite of the changes in prices of things which make up the "cost of living."

(a) The Price Quotations

The price quotations refer to commodities consumed by workingmen's families, and are taken from representative firms and districts in industrial centers. Some of the quotations are submitted to the Bureau by storekeepers, while in other cases

the Bureau's field agents collect the necessary data. The problem of keeping the identity of commodities the same is difficult, but essential uniformity is obtained by careful comparisons of grades, and by the Bureau's specifying in detail the qualities of the articles involved.

(b) The Types and Grouping of Commodities

Prices are secured for six types of commodities or services: (1) food, (2) clothing, (3) rent, (4) fuel and light, (5) furniture and furnishings, and (6) miscellaneous items.

(c) Method of Calculating the Index

The average price of each article in each group—as food, clothing, etc.—is multiplied by a weight showing the quantity of the article consumed by a family in a year. The products are then totaled. The sums give the value of all of the articles in the group at the different periods to which the prices apply. In order to get a measure of the change in the price for the group from period to period, 1913 is selected as a base, or 100 per cent, in terms of which the values for other periods are expressed as percentages. The percentage changes in each of the groups are then weighted by factors according to their relative importance in the family budget, weights being based upon the result of a study of more than 12,000 family budgets in 92 localities in the United States.[1]

(d) The Form and Place of Publication

Changes in cost of living for the country as a whole and for

[1] The group weights are as follows: food, 38.2 per cent; clothing, 16.6 per cent; rent, 13.4 per cent; fuel and light, 5.3 per cent; furniture and furnishings, 5.1 per cent; and miscellaneous, 21.3 per cent. The *National Industrial Conference Board*, New York City, publishes a similar index number of cost of living, the group weights being as follows: for food, 43.1 per cent; for shelter, 17.7 per cent; for clothing, 13.2 per cent; for fuel and light, 5.6 per cent; for sundries, 20.4 per cent. See Carr, Elma, "Cost of Living Statistics of the United States Bureau of Labor Statistics, and (of) the National Industrial Conference Board," *Journal of the American Statistical Association*, December, 1924, pp. 484-507.

different cities are published in the *Monthly Labor Review*, United States Bureau of Labor Statistics.

## 2. WHOLESALE PRICE INDEX NUMBERS ISSUED BY PRIVATE ORGANIZATIONS

A number of private organizations in the United States prepare index numbers of wholesale prices. These originally grew out of some particular need or were designed for some special purpose in connection with market analysis, special trade or financial publications, etc. While, in general, less is known about them than about the public series prepared by governmental agencies, they are widely used, quoted, and relied upon to measure price changes. Those best known are briefly described below.

### (1) *Bradstreet's Index Number*[1]

Bradstreet's wholesale index number is published monthly as a total price of 96 articles reduced to a per-pound basis.

#### a. The Price Quotations

Little is known about the source of the quotations but the compilers say they are secured from central markets.

#### b. The Types and Grouping of Commodities

The articles are divided into 13 groups as follows: (1) breadstuffs, (2) live stock, (3) provisions and groceries, (4) fresh and dried fruits, (5) hides and leather, (6) raw and manufactured textiles, (7) metals, (8) coal and coke, (9) mineral and vegetable oils, (10) naval stores, (11) building materials, (12) chemicals and drugs, and (13) miscellaneous.

[1] See "Comparison of Methods Used in Constructing Index Numbers of Wholesale Prices," *Monthly Labor Review*, September, 1920, pp. 65-70. This is a comparison of the methods used by the *Bureau of Labor Statistics*, the *Annalist*, *Bradstreet*, and *Dun*.

### c. Method of Calculating the Index

The index number for each of the thirteen groups is the su: in dollars and cents of the average price per pound of tl articles included. The index for all of the commodities is tl sum of the indexes for the groups, and the yearly numb the average of the monthly numbers. No base is used, and is not clear from the descriptions contained in *Bradstreet* whether the prices are averages of extremes or something els Moreover, the sources of the quotations are not disclosed, nc is the method described by which interpolations are made fc missing data.

Weights are not used, except as they appear in the proce of reducing all quantities to a price-per-pound basis. This, c course, results in employing a—

". . . curious combination of rational and irrational weights. Tl rational element consists in the inclusion of several quotations fc important articles like pig iron, coal, lumber, and hog product and only one quotation for articles like lemons, tea, and flax. Tl irrational element results from the reduction of all the origin quotations to prices per pound. On April 1, 1897, these prices pe pound ranged from $0.0008 for soft coal and coke to $0.52 for quicl silver and $0.83 for rubber. Recognition of the excessive influenc upon the results accorded to these high-priced articles presently le the computers to drop them from the index number; but they seer to have retained articles like alcohol and Australian wool which i 1897 cost $0.33 and $0.49 per pound—400 and 600 times as muc as soft coal and coke."[1]

### d. The Form and Place of Publication

The index is published in *Bradstreet's* both as monthly an

[1] *Bulletin of the United States Bureau of Labor Statistics,* Who Number 173, p. 101. Another writer in speaking of Bradstreet method of weighting, says, "Illogical as this system may seem, howeve it does not give the erratic results one might expect, because it is i part negatived by varying the number of commodities of each grou that is, few commodities are used in those classes of goods having hig values per pound, while many are used where value per pound is lov The Bradstreet's weighting system, then, while on its face almost ridi ulous, is not nearly so bad as it looks."

as annual numbers, the edition shortly after the beginning of each year giving a convenient review by years, months, and groups of commodities.

### (2) *Dun's Index Number*[1]

#### a. The Price Quotations

Dun's index number is based upon the wholesale prices of about 200 commodities[2] taken from the principal markets of the United States.

#### b. The Types and Grouping of Commodities

The commodities included are divided into the following groups:[3] (1) breadstuffs, (2) meats, (3) dairy and garden products, (4) other food, (5) clothing, (6) metals, (7) miscellaneous.

#### c. The Method of Calculating the Index

The index numbers are computed by (1) multiplying the price of each article by the annual per capita consumption, (2) totaling the products in each group to give the group index, and (3) totaling the group indexes to get the total index number. Concerning the method used, *Dun's Review* of May 9, 1914, says:

[1] See reference in note 1, p. 523.

[2] In a pamphlet entitled "Commodity Prices, a Record Covering a Period of Over Half a Century," taken from *Dun's Review*, January 1, 1919, it is said that "about 300 wholesale quotations are taken."

[3] "Breadstuffs include quotations of wheat, corn, oats, rye, and barley, besides beans and peas; meats include live hogs, beef, sheep, and various provisions, lard, tallow, etc.; dairy and garden include butter, eggs, vegetables and fruits; other foods include fish, condiments, sugar, rice, tobacco, etc.; clothing includes the raw material of each industry, and quotations of woolen, cotton and other textile goods, as well as hides and leatber; metals inclnde various quotations of pig iron, and partially manufactured and finished products, as well as minor metals, coal and petroleum. The miscellaneous class embraces many grades of lumber, and also lath, brick, lime, glass, turpentine, hemp, linseed oil, paints, fertilizers and drugs."—*Dun's Review*, January 10, 1925, p. 11.

"Quotations of all the necessaries of life are taken and in each case the price is multiplied by the annual per capita consumption, which precludes any one commodity having more than its proper weight in the aggregate. Thus, wide fluctuations in the price of an article little used do not materially affect the 'index,' but changes in the great staples have a large influence in advancing or depressing the total. . . . The per capita consumption used to multiply each of many hundreds of commodities does not change. There appears to be much confusion on this point, but it should be seen at a glance that there would be no accurate record of the course of prices if the ratio of consumption changed. It was possible, however, to obtain figures sufficiently accurate to give each commodity its proper importance in the compilation. This was done by taking averages for a period of years when business conditions were normal and every available trade record was utilized, in addition to official statistics of agriculture, foreign commerce, and census returns of manufactures."

The characteristics of this index number are further described by *Dun's Review* of January 10, 1925, as follows:

"It is timely to point out . . . that wholesale quotations only are used as a basis for the figures given, no attempt having been made here to measure the fluctuations in retail prices. The latter usually vary so considerably in different sections of the same city that satisfactory comparisons are difficult, if not impracticable. Nearly all barometers of price trends are based on wholesale quotations, and Dun's Index Number has the scientific foundation of making allowance for the relative importance of each of the many items that comprise the record. Obviously, some commodities enter more largely into consumption than others, and in computing an index number, a distinction should be made between a staple that is widely consumed and another article the per capita consumption of which is small. In an index number where such an allowance is not made, it follows that some articles will have a disproportionate influence upon the total, while others will not have their proper weight in the general result."

## d. The Form and Place of Publication

This number appears regularly in *Dun's Review*, New York. In the annual number, convenient summaries are given, showing price changes for commodities by groups, by months and years.

### (3) *The New York Annalist's Index Number*[1]

The *Annalist*, a New York financial journal, computes a wholesale price index number based upon 25 food products. In the issue for January 5, 1925, this number is described as showing "the food cost of living."

#### a. The Price Quotations

The quotations are taken from Chicago and New York markets and are chosen, it is claimed, so as to represent a theoretical family budget.

#### b. The Types of Commodities

The following commodities are included: steers, hogs, sheep, beef (fresh), mutton (dressed), beef (salt), pork (salt), bacon, codfish (salt), lard, potatoes, beans, flour (rye), flour (wheat, spring), flour (wheat, winter), corn meal, rice, oats, apples (evaporated), prunes, butter (creamery), butter (dairy), cheese, coffee, sugar (granulated).

#### c. The Method of Calculating the Index

The *Annalist* index number is an average of relatives, the steps in its computation being (1) to express the price of each article each period as a relative with its average price 1890-1899 as a base, (2) to sum the relatives, and (3) to take an arithmetic mean. No explicit weighting is used—the different commodities affecting the result in proportion to their relative increase or decrease as compared to the base period.[2]

#### d. The Form and Place of Publication

Weekly, monthly, and yearly numbers in the form of relatives are published currently in the weekly numbers of the journal.

[1] See reference in note 1, p. 523.
[2] See the criticism of this method, *supra*, pp. 489-493, 497.

### (4) *Professor Fisher's Index Number*

Professor Irving Fisher of Yale University publishes weekly through a syndicate of American newspapers an index number of wholesale prices in the United States, and its reciprocal the purchasing power of the dollar. The series was begun in the first week of January, 1923, a number each week from that time to date being available.

#### a. The Price Quotations

The quotations are taken from *Dun's Review*. In the beginning, 200 commodities were used; recently, however, this number has been increased to 205.

#### b. The Types of Commodities

The 205 commodities may be distributed in the following groups (the numbers in parentheses showing the percentage of the total in each group): food (45.9); clothing and cloths (16.9); paper, rubber, and fibers (2.3); metals (9.5); fuels (15.9); building materials (5.9); chemicals (3.6). Separate indexes for the groups, however, are not published.

#### c. The Method of Calculating the Index

The method of calculation is now as follows: the price of each article is multiplied by the quantity of that article sold in 1919—the United States Bureau of Labor quantities being used. The sums of the products for each week, month, or year, therefore, may be thought of as giving the total value of the articles sold at prices for the period and in quantities corresponding to those in 1919. The index numbers, however, are issued as relative or percentage numbers with 1913 as the base.[1] In this form they show "the relative value, from week

[1] In order to put the articles on the 1913 base, the 1923 series was equated on the basis of the Bureau of Labor index number (156 = 1913) for the week ending November 17, 1922. With the change in Fisher's number made in 1924, a further equating was necessary. His 1924 series is equated to his own number (151.9) for the week ending November 16, 1923.

to week, of a cargo of the 205 commodities in the above specified quantities." [1]

Previous to the 1924 revision, class weights were also used. These were chosen because it was impossible to get sufficient quotations for some of the commodities. Accordingly, correction factors or class weights were applied to the quantity weights. These, however, have been dispensed with in the 1924 revision except in the case of the chemical group. The quantity weights in this group are increased by one-half.

d. The Form and Place of Publication [2]

Fisher's series is published each Monday morning in the more important metropolitan newspapers. It appears in two forms: (1) as relative numbers based on 1913, and (2) as cents showing the purchasing power of the 1913 dollar. The second series of amounts are gotten by dividing the first series into one and multiplying by 100. That is, they are the reciprocals of the relatives.

*(5) The Commodity Price Index of Business Cycles of the Harvard Committee on Economic Research* [3]

The purpose of this index of wholesale prices is to measure changes in general business conditions. It is not intended to measure changes in the level of prices nor the effect of the changes on cost of living—the two purposes for which index numbers are generally computed.[4]

[1] Fisher, Irving, "Revision of the Weekly Index Number," *Journal of the American Statistical Association*, September, 1924, pp. 336-347 at p. 340. The reference in the quoted part is to the individual quantity weights corresponding to the different commodities.

[2] The list of commodities used in 1923 and in 1924 together with the quantity weights are shown in Fisher, *op. cit.*, pp. 341-343. This article also explains in detail the method followed including the adjustments made in 1924.

[3] This index is fully described in Persons, W. M., and Coyle, Eunice S., "A Commodity Price Index of Business Cycles," *The Review of Economic Statistics*, Preliminary Volume 3, Number II, November, 1921, pp. 353 to 369.

[4] See *supra*, pp. 480-481.

a. The Price Quotations

From an analysis of the fluctuations in the prices of a large number of commodities, 10 "varied in nature, important in industry, unusually sensitive in price, not greatly affected by the seasons, and similar with respect to their main cyclical price movements"[1] were selected.

b. The Types of Commodities

The commodities used are as follows: (1) cottonseed oil, (2) coke, (3) spelter, (4) pig iron, (5) bar iron, (6) mess pork, (7) hides, (8) print cloths, (9) sheetings, and (10) worsted yarns.

"Instead of including a large number of commodities, a few of which have great influence but most of which have little influence on the result, it is better for our purpose to include a limited number of carefully selected commodities with homogeneous cyclical price movements."[2]

c. The Method of Calculating the Index

The method of calculating the index is to take an unweighted geometric mean of the prices of the 10 commodities relative to their geometric average price in the base period, 1890-1899.

d. The Form and Place of Publication

Monthly and annual index numbers of business cycles from 1890 to September, 1921, are contained in *The Review of Economic Statistics,*[3] and current numbers in *Statistical Record,* Harvard Economic Service, Cambridge, Mass.

III. INDEX NUMBERS OF PRODUCTION

During the World War it became apparent that index num-

[1] Persons, W. M., and Coyle, Eunice S., *loc. cit.*, p. 353.
[2] *Op. cit.*, p. 356.
[3] *Loc. cit.*, p. 369.

bers of price changes did not truly represent changes in industrial and business conditions. The "dollar" became a variable rather than a fixed standard. Accordingly, the need for some measure of change in quantities of things produced, exchanged, and sold was supplied by the calculation of a number of indexes of production.

Among these indexes, those prepared by Stewart,[1] King,[2] Snyder,[3] and others were significant. Somewhat later, Professor E. E. Day, of the Harvard Committee on Economic Research, prepared quantity indexes for agriculture, manufacturing, and mining, separately and combined. The methods used in these indexes are briefly described below.

### 1. THE INDEX OF PHYSICAL PRODUCTION OF THE HARVARD COMMITTEE ON ECONOMIC RESEARCH

#### (1) *Index of Agricultural Production*[4]

##### a. Quantity Data

The annual amounts of production of twelve crops are used, the data being drawn from records of the Department of Agriculture, supplemented by similar data from other sources.

##### b. Types of Commodities

For the original index, which covered the period from 1879-1920, the annual amounts of production of the following commodities were used: hay, corn, oats, wheat, barley, rye, rice, white potatoes, sugar, tobacco, cotton, and flaxseed.

[1] Stewart, W. W., "An Index of Production," *The American Economic Review*, March, 1921, pp. 57-81.

[2] King, W. I., *Bankers' Statistics Corporation*, Special Service, Vol. 2, No. 12, August 24, 1920.

[3] Snyder, Carl (not published). See, however, *Income in the United States*, National Bureau of Economic Research, Harcourt Brace, New York, 1921, p. 79.

[4] For a detailed description of this index see Day, E. E., "An Index of the Physical Volume of Production," *The Review of Economic Statistics*, (Reprinted from the September, 1920—January, 1921, numbers, pp. 1-14).

### c. The Method of Calculating the Index

Two indexes are constructed—a so-called "unadjusted," and an "adjusted" index.

The unadjusted index is calculated as follows: (1) the quantity each year for each commodity is expressed as a relative of the amount in the base period 1909 to 1913; (2) the relatives are weighted by the average annual values of the individual crops in the same base period, 1909-1913; and (3) a weighted geometric mean is taken of the relatives.

The adjusted index is computed differently, the steps being to (1) determine the secular trend of the individual series, (2) express the original items as percentages of the ordinates of secular trend;[1] and (3) take an arithmetic mean of these percentages.

### d. The Form and Place of Publication

Both unadjusted and adjusted series for the period 1879-1920 are published by The Harvard Committee.[2]

## *(2) Index of Mining*[3]

### a. Quantity Data

The basic data for the most part are secured from the United States Geological Survey.

### b. Types of Commodities

For the original index which covered the period 1879 to 1919, the following commodities were included: gold, silver,

[1] See pp. 440, 446-447, where these terms are defined, and an illustrative example worked out.

[2] *Loc. cit.* A continuation of the unadjusted and adjusted indexes of agricultural production—certain modifications having been made from time to time—is contained in the *Review of Economic Statistics*, as follows: Preliminary Volume IV, No. 3, July, 1922, covering the period 1909 to 1921; Preliminary Volume V, No. 3, July, 1923, for the year 1922; Preliminary Vol. VI, No. 3, July, 1924, for the year 1923.

[3] See note 4, p. 531.

pig iron, copper, lead, zinc, anthracite coal, bituminous coal, petroleum, and coke.

c. The Method of Calculating the Indexes

Two indexes are computed: (1) an unadjusted, and (2) an adjusted index, the methods being identical with those used in securing the agricultural number.[1]

d. The Form and Place of Publication

Both the unadjusted and adjusted indexes are published by the Harvard Committee on Economic Research,[2] details being given for each of the commodities and for the group as a whole.

*(3) Index of Manufacture*[3]

a. Quantity Data

The quantity data are for thirty-three series covering the years 1899 to 1919, selection being based upon the availability of the data and their importance.[4]

b. Types and Grouping of Commodities

The thirty-three series are divided in ten groups as follows: (1) food; (2) textiles; (3) iron and steel; (4) lumber; (5) liquors; (6) chemicals; (7) stone, glass, and clay products; (8) metals, non-ferrous; (9) tobacco; and (10) vehicles.

[1] See above, p. 532.
[2] *Loc. cit.*, Sep., 1920—Jan., 1921, pp. 15-27. Both types of indexes covering other years are continued in the *Review of Economic Statistics*, as follows: Preliminary Vol. IV, No. 3, July, 1922, covering the years 1909 to 1921; Preliminary Vol. V., No. 3, July, 1923, covering the year 1922; and in Preliminary Vol. VI, July, 1924, for the year 1923.
[3] See note 4, p. 531.
[4] An analysis of over 80 series for the Census years 1899, 1904, 1909, and 1914, in part supplied the basis for the selection of the 33 series used in the annual index.

### c. The Method of Calculating the Indexes

The steps in the calculation of the unadjusted index are as follows: (1) computing relatives for each of the thirty-three series for each year in terms of the corresponding items in the base year, 1909; (2) applying weights to the relatives in each series based upon census data for 1909—the individual series, and the groups being separately weighted; (3) adjusting the group indexes so as to conform to those secured from a similar analysis of the census years;[1] and (4) computing a weighted geometric mean of the group indexes—the weights for the groups being the values added by manufacture as reported by the United States Census Bureau.

The adjusted index is calculated as follows: (1) determine for each of the 33 series the line of secular trend by the least-square method, the period to which the line is fitted being 1899 to 1913; (2) express the original items year by year as percentages of trend, (3) apply weights as in step two for the unadjusted index, and (4) take a weighted arithmetic average of the group indexes, the weights being the values added by manufacture, as in the unadjusted series.

### d. Form and Place of Publication

Both adjusted and unadjusted indexes are published by the Harvard Committee on Economic Research,[2] the detail covering the years 1899 to 1919.

### *(4) Combined Index of Agriculture, Mining, and Manufacture*[3]

The quantity data, and the types and grouping of commodi-

[1] See *loc. cit.*, pp. 51 and 54.

[2] *Loc. cit.*, September, 1920—January, 1921, pp. 44-63. Similar indexes for later periods are given in the following numbers of the *Review of Economic Statistics:* Preliminary Vol. V, No. 1, January, 1923, pp. 30-60, covering monthly and annual indexes, 1919 to 1922; Preliminary Vol. V., No. 3, July, 1923, pp. 205-211, Preliminary Vol. VI, No. 3, July, 1924, pp. 199-204.

[3] See note 4, p. 531.

ties are the same as those indicated above under the three separate indexes. The method of calculating a composite of the three is as follows:

The combined unadjusted index is secured by calculating each year a weighted geometric mean of the three indexes, the weights for each index being the aggregate value of production in the respective fields during the census year 1909.[1]

The combined adjusted index is a weighted arithmetic mean of the three separate indexes, the weights being the same as in the unadjusted index.[1]

## 2. OTHER INDEXES OF PHYSICAL PRODUCTION

### (1) *The Federal Reserve Board*

The Federal Reserve Board prepares and publishes each month "Indexes of Industrial Activity."[2]

### (2) *The Department of Commerce*

The United States Department of Commerce in its monthly *Survey of Current Business* publishes the following indexes:

a. "A Monthly Index of Manufacturing Production."[3]
b. "A Monthly Index of Raw Material Production."[4]
c. "A Monthly Index of Mineral Production."[5]
d. "A Monthly Index of Forestry Production."[6]

[1] *Loc. cit.*, pp. 64-68.

[2] These indexes were first presented, together with a description of data and methods, in the *Federal Reserve Bulletin*, March, 1922. A revision was made in March, 1924, the method being described in *Bulletin*, March, 1924, pp. 183-188.

[3] See *Survey of Current Business*, January, 1923, pp. 22-28, for a description of the contents of this index and the method by which it is calculated.

[4] *Ibid.*, Sept., 1922, pp. 22-24.

[5] *Ibid.*, May, 1922, pp. 19-22.

[6] *Ibid.*, August, 1922, pp. 18-21.

## IV. Indexes of Volume of Trade

### 1. "PERSONS'" INDEX OF THE HARVARD COMMITTEE ON ECONOMIC RESEARCH

"An Index of Trade for the United States"[1] for the years 1903-1923 is of a somewhat different type from those which have been termed production indexes, or price indexes designed to measure cyclical fluctuations in business.[2] The object in this case is to so combine series, such as bank clearings outside New York, values of imports of merchandise, gross earnings of railroads, production of pig iron, and the relative number of wage earners employed in industrial establishments, that the resulting index will "be responsive to variations in the general physical volume of trade."[3] The manner in which this is done is interesting but too detailed to be outlined in this place.

### 2. "SNYDER'S" NEW INDEX OF THE VOLUME OF TRADE[4]

This index is a weighted composite of 56 different series of monthly data grouped into 28 major classes, covering, among other things, productive activity; primary distribution, such as car loadings, wholesale trade, exports and imports, etc.; distribution to consumers, such as department store sales, chain store sales, mail order sales, etc.; general business activity, including shares sold on the New York stock exchange, new corporate financing, etc.

All of these various series, comprising the "immensely greater part of the nation's trade, probably 80 per cent and more,"[5] are combined into a single index in the belief that

[1] Persons, W. M., *Review of Economic Statistics*, Preliminary Vol. V, No. 2, April, 1923, pp. 71-78.

[2] See the description of the Harvard Ten Commodity Index, *supra*, pp. 529-530.

[3] Persons, W. M., *loc. cit.*, p. 78.

[4] Fully described in the *Journal of the American Statistical Association*, December, 1923, pp. 949-963.

[5] *Ibid.*, p. 950.

together they represent an index of trade far better than does the production of basic commodities alone. The different series are reduced to a common denominator in terms of their normal growth, seasonal variation, where important, being allowed for, and price changes eliminated. The index is computed as percentages of normal trend.[1]

### 3. OTHER TRADE INDEXES

#### (1) *The Federal Reserve Board*

a. "An Index of the Trend of Retail Trade."[2]
b. "An Index of Wholesale Trade."[3]

#### (2) *The United States Department of Commerce*

a. "Monthly Index of Crop Marketings."[4]
b. "Monthly Index of Marketing of Animal Products."[5]

## V. INDEXES OF GENERAL BUSINESS CONDITIONS

The foregoing indexes for the most part relate to specific phenomena, such as prices; production, including agriculture, mining, manufacturing; trade; marketing; etc. They are not designed to serve as barometers or as forecasters of business change through periods of depression, recovery, prosperity, financial strain and crisis. That is, they have to do not so much with defining and with timing the period of these shifts in business as they do with measuring on a relative basis the changes which take place.

But there is another type of index which remains to be

[1] For details see note 4, p. 536.
[2] For the method used in constructing this index, see the *Federal Reserve Bulletin*, January, 1924, pp. 17-19.
[3] *Ibid.*, April, 1923, pp. 439-442.
[4] For the method used in computing this index, see the *Survey of Current Business*, July, 1922, pp. 17-21.
[5] *Ibid.*, June, 1922, pp. 18-21.

described. It has to do with a measurement of general business conditions, and with a forecast of what they are likely to be in the future.

Certain aspects of the business cycle, so-called, were described in Chapter XIV as a background for the special treatment of time series. Something more, however, needs to be said about it.

Business conditions are always in a state of flux: they are never "normal" in the sense of being stationary. But the changes through which they pass are not fortuitous or haphazard. This has been demonstrated beyond question of doubt. Neither are they perfectly regular and periodic. The ups and downs of business do have characteristic features, however, and probably do not vary more than a few per cent [1] from what may be termed normal. Business in general, and certain of its specific phenomena pass through well-defined major and minor movements. Accordingly, it is possible to determine their order and the relations between them, to set up a measure of present conditions, and to give a forecast of what those in the immediate future are likely to be. This is what is done by the Harvard Committee on Economic Research, for instance, in its "Index of General Business Conditions," described below.

## 1. THE HARVARD INDEX OF GENERAL BUSINESS CONDITIONS

In the December, 1916, number of the *American Economic Review*,[2] Professor Warren M. Persons published a significant article. By the use of the correlation coefficient, he established the time fluctuations between a large number of series of business data and sorted out certain series which he called "a business barometer." Certain other series he found had fore-

[1] Snyder claims not more than 5 per cent plus or minus from "normal."

[2] Persons, W. M., "The Construction of a Business Barometer Based Upon Annual Data," *American Economic Review*, December, 1916, pp. 739-769.

casting properties. With this contribution as a beginning, the Harvard Committee on Economic Research now publishes weekly, as a part of its *Economic Service*, an "Index of Business Conditions."

Its barometer and forecaster are not based upon the theory that the cycles of business are perfectly periodic, nor upon the assumption that "for every action in business there is necessarily an equal and opposite reaction." They are rather founded upon the results of an elaborate study of data through the period 1903 to 1914 which showed that there is a "sequence in movements in the speculative, business, and money markets which can be measured statistically, and shown graphically on an index chart." [1]

The chart covering the trial period, 1903 to 1914, is shown in Figure 88.

An inspection of this chart shows the following important relations: (1) an interval of several months between the movements in the curves of speculation, of business and of money; and (2) the same order in the upward and downward movements and turning points of the curves. The movements are as follows: those in Curve "A" precede from six to ten months those in Curve "B"; those in Curve "B" precede from two to eight months those in Curve "C." "It is the regularity in the sequence of the movements of the three curves which affords a logical basis for scientific business forecasting. Curve 'A' moves first, 'B' second, 'C' third—speculation, business, money." [2]

The interpretation of this index is based upon "(1) the direction of the movement of each curve in relation to the movements of the other curves; (2) the direction of the immediately preceding movement; (3) the magnitude of such movements." [3]

[1] "The Harvard Index of General Business Conditions—Its Interpretation," *Harvard University Committee on Economic Research*, Cambridge, Mass., 1923, p. 8.

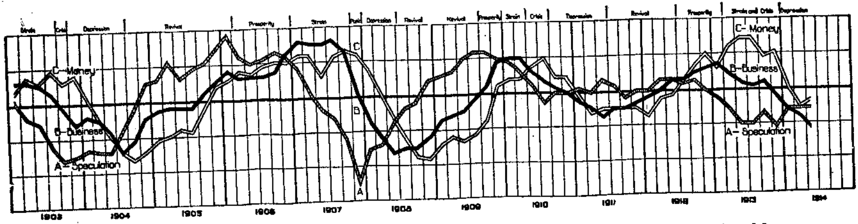
[2] *Op. cit.*, p. 9.

[3] *Ibid.*, p. 13.

FIGURE 88

The Harvard Index of General Business Conditions, 1903 to 1914 *

THE TEST-PERIOD INDEX, 1903–14




**Curve A — Speculation**
New York City bank clearings, prices of securities, averaged.

**Curve B — Business**
Wholesale commodity prices, bank clearings outside of New York City, pig-iron production, averaged.

**Curve C — Money**
Interest rates on commercial paper, loans and deposits of New York City banks, averaged.

* Reproduced by the courtesy of the *Editors* of the *Harvard Economic Review.*

The index was published in the form shown in Figure 88 until May 19, 1923. Since this date a similar chart—see Figure 89—has been published currently in the *Harvard Economic Service.*[1]

Curve "A"—speculation—is now based upon New York bank debits and industrial stock prices; Curve "B"—business—upon outside (New York City) bank debits and commodity prices; Curve "C"—money—upon interest rates on 4-6 months good, and 4-6 months prime commercial paper. While the new curves are based upon somewhat different data from the old ones, they have the same function, and their movements are to be interpreted in the same way as before the change was made.

## 2. THE BROOKMIRE FORECASTING COMPOSITE LINE[2]

The forecasting line prepared by the *Brookmire Economic Service* is not designed to show the state of business, but rather to forecast stock and commodity prices. It is made from a simple average of the following six series, all of which are treated for seasonal variation and some of them for secular trend: (1) the prices of 40 industrial and railroad stocks on the New York exchange, multiplied by the number of shares sold on this exchange; (2) a variety of series indicative of physical production; (3) the ratio of the value of imports to the value of exports; (4) the turnover of bank deposits; (5) interest rates on 4-6 months' commercial paper; and (6) the open market rate for three months' bills in London.

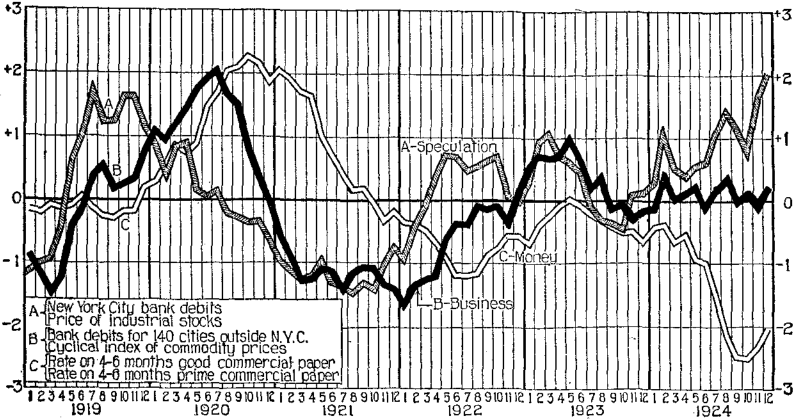
Averages for current months are compared with those for 1904-1913, the relative fluctuations being expressed in terms of the maximum. The amounts are plotted on semi-logarithmic

[1] See Persons, W. M., "The Revised Index of General Business Conditions," *The Review of Economic Statistics*, July, 1923, pp. 187-195, for an account of the necessity for revision, and the method of accomplishing it.

[2] See Vance, Ray, *Business and Investment Forecasting*, The Brookmire Economic Service, New York, 1922, for a description of the method of computing the forecasting line.

FIGURE 89

THE HARVARD INDEX CHART OF GENERAL BUSINESS CONDITIONS, 1919—1924 *




* Reproduced by the courtesy of the *Editors* of the *Harvard Economic Review.*

paper, which has the effect of toning down the extreme variations.

The direction in which the line is drawn from month to month depends upon the size of the average of the six factors as compared with that for 1904 to 1913. When the average is within a neutral zone of about 3 per cent above and below the base line, the change recorded by it is held to have no significance—the new point on the forecasting line being moved horizontally. When the average is out of the zone, it is held to be significant for forecasting purposes. If, however, within four months it crosses the neutral zone again, the whole movement is disregarded.

The forecasting curve is held to anticipate by one month changes in stock prices, and by six to seven months, changes in commodity prices.

### 3. OTHER BAROMETRIC AND FORECASTING INDEXES

Space is not available in which to describe the following indexes: the "Compositplot" of the *Babson Statistical Organization;*[1] the "money," the "stock price," and the "business" curves of the *Standard Statistics Corporation;*[1] nor the *Annalist's* "Barometer and Business Index Line."[2] The reader, however, will find a study of the "services" of these and other organizations of interest.

## VI. OTHER INDEXES OF BUSINESS AND ECONOMIC PHENOMENA

Business and statistical literature are filled with "indexes" of various types. It is inadvisable, however, in this place to do more than mention some of those which are outstanding. This is done in bibliographical form below.

[1] See Knauth, Oswald W., "Statistical Indexes of Business Conditions and Their Uses," in *Business Cycles and Unemployment*, McGraw-Hill, New York, 1923, pp. 364-368.

[2] Explained in *The Annalist*, March 28 and October 24, 1921.

### 1. MONEY AND PRICES

"An Index Chart Based on Price and Money Rates." [1]
"Index of the General Price Level." [2]
"Index of Velocity of Bank Deposits." [3]
"A New Clearings Index of Business for Fifty Years." [4]
"International Price Indexes." [5]

### 2. EMPLOYMENT AND UNEMPLOYMENT

"Index of Employment in Manufacturing Industries." [6]
"An Index of the Labor Market." [7]
"Employment and the Business Cycle." [8]
"Fluctuations of Employment in Cities of the United States, 1902—1917." [9]

[1] Persons, W. M., *Review of Economic Statistics*, January, 1922, pp. 7-11.

[2] Snyder, Carl, *Journal of the American Statistical Association*, June, 1924, pp. 189-195.

[3] Described by Burgess, W. Randolph, *Journal of the American Statistical Association*, June, 1923, pp. 727-740; *Federal Reserve Bulletin*, May, 1923; compared with Snyder's Volume of Trade Index by Snyder, Carl, "A New Index of Business Activity," *Journal of the American Statistical Association*, March, 1924, pp. 36-41.

[4] Snyder, Carl, *Journal of the American Statistical Association*, September, 1924, pp. 329-335.

[5] See *Federal Reserve Bulletin*, February, 1922, pp. 147-153; July, 1922, pp. 801-806; August, 1922, pp. 922-929; September, 1922, pp. 1052-1059. See also Snodgrass, Katharine, "A New Price Index for Great Britain," *Journal of the American Statistical Association*, June, 1922, pp. 241-249.

[6] *Federal Reserve Bulletin*. December, 1923, pp. 1272-1279. The *method of preparing this index was planned, and its construction supervised by* Professor W. A. Berridge. See also Berridge, W. A., "Cycles of Unemployment in the United States," Houghton Mifflin, Boston, 1923, for an account of the uses of such an index.

[7] *Federal Reserve Bulletin*, February, 1924, pp. 83-87. This index was planned by Dr. Berridge, Brown University.

[8] Berridge, W. A., *Review of Economic Statistics*, January, 1922, pp. 12-51. Also similar articles by the same author in *Journal of the American Statistical Association*, March, 1922, pp. 42-55; and June, 1922, pp. 227-240.

[9] Hart, Hornell, "Employment Fluctuations in the United States 1902-1917," *Studies of the Helen S. Trounstine Foundation*, Cincinnati, 1918, Vol. I, pp. 47-59.

"An Index of Factory Employment in Illinois."[1]
"An Index of the Number of Applicants per One-hundred Positions Open at Illinois Free Employment Offices."[2]

### 3. INDEX OF FOREIGN EXCHANGE RATES[3]

### 4. INDEXES OF DISTRIBUTION

"Department Store Stocks."[4]
"Department Store Sales."[5]

### 5. INDEXES OF SECURITY PRICES[6]

"An Index of Industrial Stock Prices."[7]
"A Monthly Index of Bond Yields, 1919-1923."[8]

### 6. INDEXES OF EARNINGS AND WAGE-RATES

Index numbers of trends of hourly wage-rates, weekly wage-rates, and weekly hours, are published by the United States Bureau of Labor Statistics. The methods used are described by the Bureau as follows:

"In computing the index numbers for a trade, the first step is to obtain the average rate for the trade, which is done by multiply-

[1] Method of construction described in *Annual Report* Illinois Department of Labor, 1923, Springfield, Ill. Current data appear in *The Labor Bulletin*, issued monthly by Illinois Department of Labor, Chicago, Ill.

[2] *Ibid.*

[3] See *Federal Reserve Bulletin*, July, 1921, pp. 794-799; see also a criticism of this index by Davis, J. S., "Index Numbers of Foreign Exchange," *Quarterly Journal of Economics*, May, 1922, pp. 535-542; and a reply by Goldenweiser, E. A., in *Quarterly Journal of Economics*, November, 1922, pp. 191-195.

[4] *Federal Reserve Bulletin*, March, 1924, pp. 189-190.

[5] *Ibid.*, January, 1924, pp. 17-21.

[6] For a comprehensive discussion of the problem of constructing index numbers of stock prices, see Mitchell, W. C., "A Critique of Index Numbers of Prices of Stocks," *Journal of Political Economy*, July, 1916, pp. 625-693.

[7] Frickey, Edwin, *Review of Economic Statistics*, August, 1921, pp. 264-277.

[8] Maxwell, F. W., and Matthews, A. M., *Ibid.*, July, 1923, pp. 212-217.

ing the rate per hour in each city by the number of union members in the city, adding the products, and dividing by the aggregate number of union members in the country entering into the total. These averages are brought into comparison with the average for the base year to determine the index number for each year. Grand average hourly rate, full-time weekly earnings, and weekly hours for all trades combined are obtained in the same manner as the corresponding figures were obtained for each of the several trades."[1]

"Course of Average Weekly Earnings in New York State Factories—An Index."[2]

"Index Numbers for the Wages of Common Labor."[3]

## VII. Conclusion

It is hoped that this chapter is more than informative. To know even in detail the methods which are used in computing different index numbers is of little importance if in their use the principles underlying the methods are ignored or forgotten.

The need for index numbers of various kinds, constructed according to different patterns helps partly but not wholly to explain the variety of types available. Too frequently, in the past, methods were followed because they were simple rather than because they were appropriate. So long as information was lacking as to the relations between methods and results, there was some justification for this condition. But that time has passed. There is no excuse to-day for the mistaken belief that all index numbers are equally good, and that from those available relating to prices, trade, unemployment, etc., selection may be made at random in order to measure business, social, and industrial changes.

[1] "Methods of Procuring and Computing Statistical Information of the Bureau of Labor Statistics," *Bulletin* 326, *United States Bureau of Labor Statistics*, Washington, D. C., March, 1923, p. 3. Current data appear in the *Monthly Labor Review* for each year, and are cumulated in the report called *Union Scale of Wages and Hours of Labor.*

[2] Published monthly in *The Industrial Bulletin*, Industrial Commission of New York State, Albany, New York.

[3] Burgess, W. Randolph, *Journal of the American Statistical Association*, March, 1922, pp. 101-103.

# APPENDIX

TABLE 1—POWERS, ROOTS, RECIPROCALS [1]

TABLE 2—COMMON LOGARITHMS—FOUR PLACES [1]

[1] From *Logarithmic and Trigonometric Tables*, Revised Edition, edited by E. R. Hedrick. Copyright, 1920, by The Macmillan Company. Reprinted by permission of the editor and publishers.

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **1.00** | 1.0000 | 1.00000 | 3.16228 | 1.00000 | 1.00000 | 2.15443 | 4.64159 | 1.00000 |
| 1.01 | 1.0201 | 1.00499 | 3.17805 | 1.03030 | 1.00332 | 2.16159 | 4.65701 | .990099 |
| 1.02 | 1.0404 | 1.00995 | 3.19374 | 1.06121 | 1.00662 | 2.16870 | 4.67233 | .980392 |
| 1.03 | 1.0609 | 1.01489 | 3.20936 | 1.09273 | 1.00990 | 2.17577 | 4.68755 | .970874 |
| 1.04 | 1.0816 | 1.01980 | 3.22490 | 1.12486 | 1.01316 | 2.18279 | 4.70267 | .961538 |
| 1.05 | 1.1025 | 1.02470 | 3.24037 | 1.15762 | 1.01640 | 2.18976 | 4.71769 | .952381 |
| 1.06 | 1.1236 | 1.02956 | 3.25576 | 1.19102 | 1.01961 | 2.19669 | 4.73262 | .943396 |
| 1.07 | 1.1449 | 1.03441 | 3.27109 | 1.22504 | 1.02281 | 2.20358 | 4.74746 | .934579 |
| 1.08 | 1.1664 | 1.03923 | 3.28634 | 1.25971 | 1.02599 | 2.21042 | 4.76220 | .925926 |
| 1.09 | 1.1881 | 1.04403 | 3.30151 | 1.29503 | 1.02914 | 2.21722 | 4.77686 | .917431 |
| **1.10** | 1.2100 | 1.04881 | 3.31662 | 1.33100 | 1.03228 | 2.22398 | 4.79142 | .909091 |
| 1.11 | 1.2321 | 1.05357 | 3.33167 | 1.36763 | 1.03540 | 2.23070 | 4.80590 | .900901 |
| 1.12 | 1.2544 | 1.05830 | 3.34664 | 1.40493 | 1.03850 | 2.23738 | 4.82028 | .892857 |
| 1.13 | 1.2769 | 1.06301 | 3.36155 | 1.44290 | 1.04158 | 2.24402 | 4.83459 | .884956 |
| 1.14 | 1.2996 | 1.06771 | 3.37639 | 1.48154 | 1.04464 | 2.25062 | 4.84881 | .877193 |
| 1.15 | 1.3225 | 1.07238 | 3.39116 | 1.52088 | 1.04769 | 2.25718 | 4.86294 | .869565 |
| 1.16 | 1.3456 | 1.07703 | 3.40588 | 1.56090 | 1.05072 | 2.26370 | 4.87700 | .862069 |
| 1.17 | 1.3689 | 1.08167 | 3.42053 | 1.60161 | 1.05373 | 2.27019 | 4.89097 | .854701 |
| 1.18 | 1.3924 | 1.08628 | 3.43511 | 1.64303 | 1.05672 | 2.27664 | 4.90487 | .847458 |
| 1.19 | 1.4161 | 1.09087 | 3.44964 | 1.68516 | 1.05970 | 2.28305 | 4.91868 | .840336 |
| **1.20** | 1.4400 | 1.09545 | 3.46410 | 1.72800 | 1.06266 | 2.28943 | 4.93242 | .833333 |
| 1.21 | 1.4641 | 1.10000 | 3.47851 | 1.77156 | 1.06560 | 2.29577 | 4.94609 | .826446 |
| 1.22 | 1.4884 | 1.10454 | 3.49285 | 1.81585 | 1.06853 | 2.30208 | 4.95968 | .819672 |
| 1.23 | 1.5129 | 1.10905 | 3.50714 | 1.86087 | 1.07144 | 2.30835 | 4.97319 | .813008 |
| 1.24 | 1.5376 | 1.11355 | 3.52136 | 1.90662 | 1.07434 | 2.31459 | 4.98663 | .806452 |
| 1.25 | 1.5625 | 1.11803 | 3.53553 | 1.95312 | 1.07722 | 2.32079 | 5.00000 | .800000 |
| 1.26 | 1.5876 | 1.12250 | 3.54965 | 2.00038 | 1.08008 | 2.32697 | 5.01330 | .793651 |
| 1.27 | 1.6129 | 1.12694 | 3.56371 | 2.04838 | 1.08293 | 2.33311 | 5.02653 | .787402 |
| 1.28 | 1.6384 | 1.13137 | 3.57771 | 2.09715 | 1.08577 | 2.33921 | 5.03968 | .781250 |
| 1.29 | 1.6641 | 1.13578 | 3.59166 | 2.14669 | 1.08859 | 2.34529 | 5.05277 | .775194 |
| **1.30** | 1.6900 | 1.14018 | 3.60555 | 2.19700 | 1.09139 | 2.35133 | 5.06580 | .769231 |
| 1.31 | 1.7161 | 1.14455 | 3.61939 | 2.24809 | 1.09418 | 2.35735 | 5.07875 | .763359 |
| 1.32 | 1.7424 | 1.14891 | 3.63318 | 2.29997 | 1.09696 | 2.36333 | 5.09164 | .757576 |
| 1.33 | 1.7689 | 1.15326 | 3.64692 | 2.35264 | 1.09972 | 2.36928 | 5.10447 | .751880 |
| 1.34 | 1.7956 | 1.15758 | 3.66060 | 2.40610 | 1.10247 | 2.37521 | 5.11723 | .746269 |
| 1.35 | 1.8225 | 1.16190 | 3.67423 | 2.46038 | 1.10521 | 2.38110 | 5.12993 | .740741 |
| 1.36 | 1.8496 | 1.16619 | 3.68782 | 2.51546 | 1.10793 | 2.38697 | 5.14256 | .735294 |
| 1.37 | 1.8769 | 1.17047 | 3.70135 | 2.57135 | 1.11064 | 2.39280 | 5.15514 | .729927 |
| 1.38 | 1.9044 | 1.17473 | 3.71484 | 2.62807 | 1.11334 | 2.39861 | 5.16765 | .724638 |
| 1.39 | 1.9321 | 1.17898 | 3.72827 | 2.68562 | 1.11602 | 2.40439 | 5.18010 | .719424 |
| **1.40** | 1.9600 | 1.18322 | 3.74166 | 2.74400 | 1.11869 | 2.41014 | 5.19249 | .714286 |
| 1.41 | 1.9881 | 1.18743 | 3.75500 | 2.80322 | 1.12135 | 2.41587 | 5.20483 | .709220 |
| 1.42 | 2.0164 | 1.19164 | 3.76829 | 2.86329 | 1.12399 | 2.42156 | 5.21710 | .704225 |
| 1.43 | 2.0449 | 1.19583 | 3.78153 | 2.92421 | 1.12662 | 2.42724 | 5.22932 | .699301 |
| 1.44 | 2.0736 | 1.20000 | 3.79473 | 2.98598 | 1.12924 | 2.43288 | 5.24148 | .694444 |
| 1.45 | 2.1025 | 1.20416 | 3.80789 | 3.04862 | 1.13185 | 2.43850 | 5.25359 | .689655 |
| 1.46 | 2.1316 | 1.20830 | 3.82099 | 3.11214 | 1.13445 | 2.44409 | 5.26564 | .684932 |
| 1.47 | 2.1609 | 1.21244 | 3.83406 | 3.17652 | 1.13703 | 2.44966 | 5.27763 | .680272 |
| 1.48 | 2.1904 | 1.21655 | 3.84708 | 3.24179 | 1.13960 | 2.45520 | 5.28957 | .675676 |
| 1.49 | 2.2201 | 1.22066 | 3.86005 | 3.30795 | 1.14216 | 2.46072 | 5.30146 | .671141 |
| **1.50** | 2.2500 | 1.22474 | 3.87298 | 3.37500 | 1.14471 | 2.46621 | 5.31329 | .666667 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **1.50** | 2.2500 | 1.22474 | 3.87298 | 3.37500 | 1.14471 | 2.46621 | 5.31329 | .666667 |
| 1.51 | 2.2801 | 1.22882 | 3.88587 | 3.44295 | 1.14725 | 2.47168 | 5.32507 | .662252 |
| 1.52 | 2.3104 | 1.23288 | 3.89872 | 3.51181 | 1.14978 | 2.47712 | 5.33680 | .657895 |
| 1.53 | 2.3409 | 1.23693 | 3.91152 | 3.58158 | 1.15230 | 2.48255 | 5.34848 | .653595 |
| 1.54 | 2.3716 | 1.24097 | 3.92428 | 3.65226 | 1.15480 | 2.48794 | 5.36011 | .649351 |
| 1.55 | 2.4025 | 1.24499 | 3.93700 | 3.72388 | 1.15729 | 2.49332 | 5.37169 | .645161 |
| 1.56 | 2.4336 | 1.24900 | 3.94968 | 3.79642 | 1.15978 | 2.49867 | 5.38321 | .641026 |
| 1.57 | 2.4649 | 1.25300 | 3.96232 | 3.86989 | 1.16225 | 2.50399 | 5.39469 | .636943 |
| 1.58 | 2.4964 | 1.25698 | 3.97492 | 3.94431 | 1.16471 | 2.50930 | 5.40612 | .632911 |
| 1.59 | 2.5281 | 1.26095 | 3.98748 | 4.01968 | 1.16717 | 2.51458 | 5.41750 | .628931 |
| **1.60** | 2.5600 | 1.26491 | 4.00000 | 4.09600 | 1.16961 | 2.51984 | 5.42884 | .625000 |
| 1.61 | 2.5921 | 1.26886 | 4.01248 | 4.17328 | 1.17204 | 2.52508 | 5.44012 | .621118 |
| 1.62 | 2.6244 | 1.27279 | 4.02492 | 4.25153 | 1.17446 | 2.53030 | 5.45136 | .617284 |
| 1.63 | 2.6569 | 1.27671 | 4.03733 | 4.33075 | 1.17687 | 2.53549 | 5.46256 | .613497 |
| 1.64 | 2.6896 | 1.28062 | 4.04969 | 4.41094 | 1.17927 | 2.54067 | 5.47370 | .609756 |
| 1.65 | 2.7225 | 1.28452 | 4.06202 | 4.49212 | 1.18167 | 2.54582 | 5.48481 | .606061 |
| 1.66 | 2.7556 | 1.28841 | 4.07431 | 4.57430 | 1.18405 | 2.55095 | 5.49586 | .602410 |
| 1.67 | 2.7889 | 1.29228 | 4.08656 | 4.65746 | 1.18642 | 2.55607 | 5.50688 | .598802 |
| 1.68 | 2.8224 | 1.29615 | 4.09878 | 4.74163 | 1.18878 | 2.56116 | 5.51785 | .595238 |
| 1.69 | 2.8561 | 1.30000 | 4.11096 | 4.82681 | 1.19114 | 2.56623 | 5.52877 | .591716 |
| **1.70** | 2.8900 | 1.30384 | 4.12311 | 4.91300 | 1.19348 | 2.57128 | 5.53966 | .588235 |
| 1.71 | 2.9241 | 1.30767 | 4.13521 | 5.00021 | 1.19582 | 2.57631 | 5.55050 | .584795 |
| 1.72 | 2.9584 | 1.31149 | 4.14729 | 5.08845 | 1.19815 | 2.58133 | 5.56130 | .581395 |
| 1.73 | 2.9929 | 1.31529 | 4.15933 | 5.17772 | 1.20046 | 2.58632 | 5.57205 | .578035 |
| 1.74 | 3.0276 | 1.31909 | 4.17133 | 5.26802 | 1.20277 | 2.59129 | 5.58277 | .574713 |
| 1.75 | 3.0625 | 1.32288 | 4.18330 | 5.35938 | 1.20507 | 2.59625 | 5.59344 | .571429 |
| 1.76 | 3.0976 | 1.32665 | 4.19524 | 5.45178 | 1.20736 | 2.60118 | 5.60408 | .568182 |
| 1.77 | 3.1329 | 1.33041 | 4.20714 | 5.54523 | 1.20964 | 2.60610 | 5.61467 | .564972 |
| 1.78 | 3.1684 | 1.33417 | 4.21900 | 5.63975 | 1.21192 | 2.61100 | 5.62523 | .561798 |
| 1.79 | 3.2041 | 1.33791 | 4.23084 | 5.73534 | 1.21418 | 2.61588 | 5.63574 | .558659 |
| **1.80** | 3.2400 | 1.34164 | 4.24264 | 5.83200 | 1.21644 | 2.62074 | 5.64622 | .555556 |
| 1.81 | 3.2761 | 1.34536 | 4.25441 | 5.92974 | 1.21869 | 2.62559 | 5.65665 | .552486 |
| 1.82 | 3.3124 | 1.34907 | 4.26615 | 6.02857 | 1.22093 | 2.63041 | 5.66705 | .549451 |
| 1.83 | 3.3489 | 1.35277 | 4.27785 | 6.12849 | 1.22316 | 2.63522 | 5.67741 | .546448 |
| 1.84 | 3.3856 | 1.35647 | 4.28952 | 6.22950 | 1.22539 | 2.64001 | 5.68773 | .543478 |
| 1.85 | 3.4225 | 1.36015 | 4.30116 | 6.33162 | 1.22760 | 2.64479 | 5.69802 | .540541 |
| 1.86 | 3.4596 | 1.36382 | 4.31277 | 6.43486 | 1.22981 | 2.64954 | 5.70827 | .537634 |
| 1.87 | 3.4969 | 1.36748 | 4.32435 | 6.53920 | 1.23201 | 2.65428 | 5.71848 | .534759 |
| 1.88 | 3.5344 | 1.37113 | 4.33590 | 6.64467 | 1.23420 | 2.65901 | 5.72865 | .531915 |
| 1.89 | 3.5721 | 1.37477 | 4.34741 | 6.75127 | 1.23639 | 2.66371 | 5.73879 | .529101 |
| **1.90** | 3.6100 | 1.37840 | 4.35890 | 6.85900 | 1.23856 | 2.66840 | 5.74890 | .526316 |
| 1.91 | 3.6481 | 1.38203 | 4.37035 | 6.96787 | 1.24073 | 2.67307 | 5.75897 | .523560 |
| 1.92 | 3.6864 | 1.38564 | 4.38178 | 7.07789 | 1.24289 | 2.67773 | 5.76900 | .520833 |
| 1.93 | 3.7249 | 1.38924 | 4.39318 | 7.18906 | 1.24505 | 2.68237 | 5.77900 | .518135 |
| 1.94 | 3.7636 | 1.39284 | 4.40454 | 7.30138 | 1.24719 | 2.68700 | 5.78896 | .515464 |
| 1.95 | 3.8025 | 1.39642 | 4.41588 | 7.41488 | 1.24933 | 2.69161 | 5.79889 | .512821 |
| 1.96 | 3.8416 | 1.40000 | 4.42719 | 7.52954 | 1.25146 | 2.69620 | 5.80879 | .510204 |
| 1.97 | 3.8809 | 1.40357 | 4.43847 | 7.64537 | 1.25359 | 2.70078 | 5.81865 | .507614 |
| 1.98 | 3.9204 | 1.40712 | 4.44972 | 7.76239 | 1.25571 | 2.70534 | 5.82848 | .505051 |
| 1.99 | 3.9601 | 1.41067 | 4.46094 | 7.88060 | 1.25782 | 2.70989 | 5.83827 | .502513 |
| **2.00** | 4.0000 | 1.41421 | 4.47214 | 8.00000 | 1.25992 | 2.71442 | 5.84804 | .500000 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **2.00** | 4.0000 | 1.41421 | 4.47214 | 8.00000 | 1.25992 | 2.71442 | 5.84804 | .500000 |
| 2.01 | 4.0401 | 1.41774 | 4.48330 | 8.12060 | 1.26202 | 2.71893 | 5.85777 | .497512 |
| 2.02 | 4.0804 | 1.42127 | 4.49444 | 8.24241 | 1.26411 | 2.72344 | 5.86746 | .495050 |
| 2.03 | 4.1209 | 1.42478 | 4.50555 | 8.36543 | 1.26619 | 2.72792 | 5.87713 | .492611 |
| 2.04 | 4.1616 | 1.42829 | 4.51664 | 8.48966 | 1.26827 | 2.73239 | 5.88677 | .490196 |
| 2.05 | 4.2025 | 1.43178 | 4.52769 | 8.61512 | 1.27033 | 2.73685 | 5.89637 | .487805 |
| 2.06 | 4.2436 | 1.43527 | 4.53872 | 8.74182 | 1.27240 | 2.74129 | 5.90594 | .485437 |
| 2.07 | 4.2849 | 1.43875 | 4.54973 | 8.86974 | 1.27445 | 2.74572 | 5.91548 | .483092 |
| 2.08 | 4.3264 | 1.44222 | 4.56070 | 8.99891 | 1.27650 | 2.75014 | 5.92499 | .480769 |
| 2.09 | 4.3681 | 1.44568 | 4.57165 | 9.12933 | 1.27854 | 2.75454 | 5.93447 | .478469 |
| **2.10** | 4.4100 | 1.44914 | 4.58258 | 9.26100 | 1.28058 | 2.75892 | 5.94392 | .476190 |
| 2.11 | 4.4521 | 1.45258 | 4.59347 | 9.39393 | 1.28261 | 2.76330 | 5.95334 | .473934 |
| 2.12 | 4.4944 | 1.45602 | 4.60435 | 9.52813 | 1.28463 | 2.76766 | 5.96273 | .471698 |
| 2.13 | 4.5369 | 1.45945 | 4.61519 | 9.66360 | 1.28665 | 2.77200 | 5.97209 | .469434 |
| 2.14 | 4.5796 | 1.46287 | 4.62601 | 9.80034 | 1.28866 | 2.77633 | 5.98142 | .467290 |
| 2.15 | 4.6225 | 1.46629 | 4.63681 | 9.93838 | 1.29066 | 2.78065 | 5.99073 | .465116 |
| 2.16 | 4.6656 | 1.46969 | 4.64758 | 10.0777 | 1.29266 | 2.78495 | 6.00000 | .462963 |
| 2.17 | 4.7089 | 1.47309 | 4.65833 | 10.2183 | 1.29465 | 2.78924 | 6.00925 | .460829 |
| 2.18 | 4.7524 | 1.47648 | 4.66905 | 10.3602 | 1.29664 | 2.79352 | 6.01846 | .458716 |
| 2.19 | 4.7961 | 1.47986 | 4.67974 | 10.5035 | 1.29862 | 2.79779 | 6.02765 | .456621 |
| **2.20** | 4.8400 | 1.48324 | 4.69042 | 10.6480 | 1.30059 | 2.80204 | 6.03681 | .454545 |
| 2.21 | 4.8841 | 1.48661 | 4.70106 | 10.7939 | 1.30256 | 2.80628 | 6.04594 | .452489 |
| 2.22 | 4.9284 | 1.48997 | 4.71169 | 10.9410 | 1.30452 | 2.81050 | 6.05505 | .450450 |
| 2.23 | 4.9729 | 1.49332 | 4.72229 | 11.0896 | 1.30648 | 2.81472 | 6.06413 | .448430 |
| 2.24 | 5.0176 | 1.49666 | 4.73286 | 11.2394 | 1.30843 | 2.81892 | 6.07318 | .446429 |
| 2.25 | 5.0625 | 1.50000 | 4.74342 | 11.3906 | 1.31037 | 2.82311 | 6.08220 | .444444 |
| 2.26 | 5.1076 | 1.50333 | 4.75395 | 11.5432 | 1.31231 | 2.82728 | 6.09120 | .442478 |
| 2.27 | 5.1529 | 1.50665 | 4.76445 | 11.6971 | 1.31424 | 2.83145 | 6.10017 | .440529 |
| 2.28 | 5.1984 | 1.50997 | 4.77493 | 11.8524 | 1.31617 | 2.83560 | 6.10911 | .438596 |
| 2.29 | 5.2441 | 1.51327 | 4.78539 | 12.0090 | 1.31809 | 2.83974 | 6.11803 | .436681 |
| **2.30** | 5.2900 | 1.51658 | 4.79583 | 12.1670 | 1.32001 | 2.84387 | 6.12693 | .434783 |
| 2.31 | 5.3361 | 1.51987 | 4.80625 | 12.3264 | 1.32192 | 2.84798 | 6.13579 | .432900 |
| 2.32 | 5.3824 | 1.52315 | 4.81664 | 12.4872 | 1.32382 | 2.85209 | 6.14463 | .431034 |
| 2.33 | 5.4289 | 1.52643 | 4.82701 | 12.6493 | 1.32572 | 2.85618 | 6.15345 | .429185 |
| 2.34 | 5.4756 | 1.52971 | 4.83735 | 12.8129 | 1.32761 | 2.86026 | 6.16224 | .427350 |
| 2.35 | 5.5225 | 1.53297 | 4.84768 | 12.9779 | 1.32950 | 2.86433 | 6.17101 | .425532 |
| 2.36 | 5.5696 | 1.53623 | 4.85798 | 13.1443 | 1.33139 | 2.86838 | 6.17975 | .423729 |
| 2.37 | 5.6169 | 1.53948 | 4.86826 | 13.3121 | 1.33326 | 2.87243 | 6.18846 | .421941 |
| 2.38 | 5.6644 | 1.54272 | 4.87852 | 13.4813 | 1.33514 | 2.87646 | 6.19715 | .420168 |
| 2.39 | 5.7121 | 1.54596 | 4.88876 | 13.6519 | 1.33700 | 2.88049 | 6.20582 | .418410 |
| **2.40** | 5.7600 | 1.54919 | 4.89898 | 13.8240 | 1.33887 | 2.88450 | 6.21447 | .416667 |
| 2.41 | 5.8081 | 1.55242 | 4.90918 | 13.9975 | 1.34072 | 2.88850 | 6.22308 | .414938 |
| 2.42 | 5.8564 | 1.55563 | 4.91935 | 14.1725 | 1.34257 | 2.89249 | 6.23168 | .413223 |
| 2.43 | 5.9049 | 1.55885 | 4.92950 | 14.3489 | 1.34442 | 2.89647 | 6.24025 | .411523 |
| 2.44 | 5.9536 | 1.56205 | 4.93964 | 14.5268 | 1.34626 | 2.90044 | 6.24880 | .409836 |
| 2.45 | 6.0025 | 1.56525 | 4.94975 | 14.7061 | 1.34810 | 2.90439 | 6.25732 | .408163 |
| 2.46 | 6.0516 | 1.56844 | 4.95984 | 14.8869 | 1.34993 | 2.90834 | 6.26583 | .406504 |
| 2.47 | 6.1009 | 1.57162 | 4.96991 | 15.0692 | 1.35176 | 2.91227 | 6.27431 | .404858 |
| 2.48 | 6.1504 | 1.57480 | 4.97996 | 15.2530 | 1.35358 | 2.91620 | 6.28276 | .403226 |
| 2.49 | 6.2001 | 1.57797 | 4.98999 | 15.4382 | 1.35540 | 2.92011 | 6.29119 | .401606 |
| **2.50** | 6.2500 | 1.58114 | 5.00000 | 15.6250 | 1.35721 | 2.92402 | 6.29961 | .400000 |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **2.50** | 6.2500 | 1.58114 | 5.00000 | 15.6250 | 1.35721 | 2.92402 | 6.29961 | .400000 |
| 2.51 | 6.3001 | 1.58430 | 5.00999 | 15.8133 | 1.35902 | 2.92791 | 6.30799 | .398406 |
| 2.52 | 6.3504 | 1.58745 | 5.01996 | 16.0030 | 1.36082 | 2.93179 | 6.31636 | .396825 |
| 2.53 | 6.4009 | 1.59060 | 5.02991 | 16.1943 | 1.36262 | 2.93567 | 6.32470 | .395257 |
| 2.54 | 6.4516 | 1.59374 | 5.03984 | 16.3871 | 1.36441 | 2.93953 | 6.33303 | .393701 |
| 2.55 | 6.5025 | 1.59687 | 5.04975 | 16.5814 | 1.36620 | 2.94338 | 6.34133 | .392157 |
| 2.56 | 6.5536 | 1.60000 | 5.05964 | 16.7772 | 1.36798 | 2.94723 | 6.34960 | .390625 |
| 2.57 | 6.6049 | 1.60312 | 5.06952 | 16.9746 | 1.36976 | 2.95106 | 6.35786 | .389105 |
| 2.58 | 6.6564 | 1.60624 | 5.07937 | 17.1735 | 1.37153 | 2.95488 | 6.36610 | .387597 |
| 2.59 | 6.7081 | 1.60935 | 5.08920 | 17.3740 | 1.37330 | 2.95869 | 6.37431 | .386100 |
| **2.60** | 6.7600 | 1.61245 | 5.09902 | 17.5760 | 1.37507 | 2.96250 | 6.38250 | .384615 |
| 2.61 | 6.8121 | 1.61555 | 5.10882 | 17.7796 | 1.37683 | 2.96629 | 6.39068 | .383142 |
| 2.62 | 6.8644 | 1.61864 | 5.11859 | 17.9847 | 1.37859 | 2.97007 | 6.39883 | .381679 |
| 2.63 | 6.9169 | 1.62173 | 5.12835 | 18.1914 | 1.38034 | 2.97385 | 6.40696 | .380228 |
| 2.64 | 6.9696 | 1.62481 | 5.13809 | 18.3997 | 1.38208 | 2.97761 | 6.41507 | .378788 |
| 2.65 | 7.0225 | 1.62788 | 5.14782 | 18.6096 | 1.38383 | 2.98137 | 6.42316 | .377358 |
| 2.66 | 7.0756 | 1.63095 | 5.15752 | 18.8211 | 1.38557 | 2.98511 | 6.43123 | .375940 |
| 2.67 | 7.1289 | 1.63401 | 5.16720 | 19.0342 | 1.38730 | 2.98885 | 6.43928 | .374532 |
| 2.68 | 7.1824 | 1.63707 | 5.17687 | 19.2488 | 1.38903 | 2.99257 | 6.44731 | .373134 |
| 2.69 | 7.2361 | 1.64012 | 5.18652 | 19.4651 | 1.39076 | 2.99629 | 6.45531 | .371747 |
| **2.70** | 7.2900 | 1.64317 | 5.19615 | 19.6830 | 1.39248 | 3.00000 | 6.46330 | .370370 |
| 2.71 | 7.3441 | 1.64621 | 5.20577 | 19.9025 | 1.39419 | 3.00370 | 6.47127 | .369004 |
| 2.72 | 7.3984 | 1.64924 | 5.21536 | 20.1236 | 1.39591 | 3.00739 | 6.47922 | .367647 |
| 2.73 | 7.4529 | 1.65227 | 5.22494 | 20.3464 | 1.39761 | 3.01107 | 6.48715 | .366300 |
| 2.74 | 7.5076 | 1.65529 | 5.23450 | 20.5708 | 1.39932 | 3.01474 | 6.49507 | .364964 |
| 2.75 | 7.5625 | 1.65831 | 5.24404 | 20.7969 | 1.40102 | 3.01841 | 6.50296 | .363636 |
| 2.76 | 7.6176 | 1.66132 | 5.25357 | 21.0246 | 1.40272 | 3.02206 | 6.51083 | .362319 |
| 2.77 | 7.6729 | 1.66433 | 5.26308 | 21.2539 | 1.40441 | 3.02570 | 6.51868 | .361011 |
| 2.78 | 7.7284 | 1.66733 | 5.27257 | 21.4850 | 1.40610 | 3.02934 | 6.52652 | .359712 |
| 2.79 | 7.7841 | 1.67033 | 5.28205 | 21.7176 | 1.40778 | 3.03297 | 6.53434 | .358423 |
| **2.80** | 7.8400 | 1.67332 | 5.29150 | 21.9520 | 1.40946 | 3.03659 | 6.54213 | .357143 |
| 2.81 | 7.8961 | 1.67631 | 5.30094 | 22.1880 | 1.41114 | 3.04020 | 6.54991 | .355872 |
| 2.82 | 7.9524 | 1.67929 | 5.31037 | 22.4258 | 1.41281 | 3.04380 | 6.55767 | .354610 |
| 2.83 | 8.0089 | 1.68226 | 5.31977 | 22.6652 | 1.41448 | 3.04740 | 6.56541 | .353357 |
| 2.84 | 8.0656 | 1.68523 | 5.32917 | 22.9063 | 1.41614 | 3.05098 | 6.57314 | .352113 |
| 2.85 | 8.1225 | 1.68819 | 5.33854 | 23.1491 | 1.41780 | 3.05456 | 6.58084 | .350877 |
| 2.86 | 8.1796 | 1.69115 | 5.34790 | 23.3937 | 1.41946 | 3.05813 | 6.58853 | .349650 |
| 2.87 | 8.2369 | 1.69411 | 5.35724 | 23.6399 | 1.42111 | 3.06169 | 6.59620 | .348432 |
| 2.88 | 8.2944 | 1.69706 | 5.36656 | 23.8879 | 1.42276 | 3.06524 | 6.60385 | .347222 |
| 2.89 | 8.3521 | 1.70000 | 5.37587 | 24.1376 | 1.42440 | 3.06878 | 6.61149 | .346021 |
| **2.90** | 8.4100 | 1.70294 | 5.38516 | 24.3890 | 1.42604 | 3.07232 | 6.61911 | .344828 |
| 2.91 | 8.4681 | 1.70587 | 5.39444 | 24.6422 | 1.42768 | 3.07584 | 6.62671 | .343643 |
| 2.92 | 8.5264 | 1.70880 | 5.40370 | 24.8971 | 1.42931 | 3.07936 | 6.63429 | .342466 |
| 2.93 | 8.5849 | 1.71172 | 5.41295 | 25.1538 | 1.43094 | 3.08287 | 6.64185 | .341297 |
| 2.94 | 8.6436 | 1.71464 | 5.42218 | 25.4122 | 1.43257 | 3.08638 | 6.64940 | .340136 |
| 2.95 | 8.7025 | 1.71756 | 5.43139 | 25.6724 | 1.43419 | 3.08987 | 6.65693 | .338983 |
| 2.96 | 8.7616 | 1.72047 | 5.44059 | 25.9343 | 1.43581 | 3.09336 | 6.66444 | .337838 |
| 2.97 | 8.8209 | 1.72337 | 5.44977 | 26.1981 | 1.43743 | 3.09684 | 6.67194 | .336700 |
| 2.98 | 8.8804 | 1.72627 | 5.45894 | 26.4636 | 1.43904 | 3.10031 | 6.67942 | .335570 |
| 2.99 | 8.9401 | 1.72916 | 5.46809 | 26.7309 | 1.44065 | 3.10378 | 6.68688 | .334448 |
| **3.00** | 9.0000 | 1.73205 | 5.47723 | 27.0000 | 1.44225 | 3.10723 | 6.69433 | .333333 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **3.00** | 9.0000 | 1.73205 | 5.47723 | 27.0000 | 1.44225 | 3.10723 | 6.69433 | .333333 |
| 3.01 | 9.0601 | 1.73494 | 5.48635 | 27.2709 | 1.44385 | 3.11068 | 6.70176 | .332226 |
| 3.02 | 9.1204 | 1.73781 | 5.49545 | 27.5436 | 1.44545 | 3.11412 | 6.70917 | .331126 |
| 3.03 | 9.1809 | 1.74069 | 5.50454 | 27.8181 | 1.44704 | 3.11756 | 6.71657 | .330033 |
| 3.04 | 9.2416 | 1.74356 | 5.51362 | 28.0945 | 1.44863 | 3.12098 | 6.72395 | .328947 |
| 3.05 | 9.3025 | 1.74642 | 5.52268 | 28.3726 | 1.45022 | 3.12440 | 6.73132 | .327869 |
| 3.06 | 9.3636 | 1.74929 | 5.53173 | 28.6526 | 1.45180 | 3.12781 | 6.73866 | .326797 |
| 3.07 | 9.4249 | 1.75214 | 5.54076 | 28.9344 | 1.45338 | 3.13121 | 6.74600 | .325733 |
| 3.08 | 9.4864 | 1.75499 | 5.54977 | 29.2181 | 1.45496 | 3.13461 | 6.75331 | .324675 |
| 3.09 | 9.5481 | 1.75784 | 5.55878 | 29.5036 | 1.45653 | 3.13800 | 6.76061 | .323625 |
| **3.10** | 9.6100 | 1.76068 | 5.56776 | 29.7910 | 1.45810 | 3.14138 | 6.76790 | .322581 |
| 3.11 | 9.6721 | 1.76352 | 5.57674 | 30.0802 | 1.45967 | 3.14475 | 6.77517 | .321543 |
| 3.12 | 9.7344 | 1.76635 | 5.58570 | 30.3713 | 1.46123 | 3.14812 | 6.78242 | .320513 |
| 3.13 | 9.7969 | 1.76918 | 5.59464 | 30.6643 | 1.46279 | 3.15148 | 6.78966 | .319489 |
| 3.14 | 9.8596 | 1.77200 | 5.60357 | 30.9591 | 1.46434 | 3.15483 | 6.79688 | .318471 |
| 3.15 | 9.9225 | 1.77482 | 5.61249 | 31.2559 | 1.46590 | 3.15818 | 6.80409 | .317460 |
| 3.16 | 9.9856 | 1.77764 | 5.62139 | 31.5545 | 1.46745 | 3.16152 | 6.81128 | .316456 |
| 3.17 | 10.0489 | 1.78045 | 5.63028 | 31.8550 | 1.46899 | 3.16485 | 6.81846 | .315457 |
| 3.18 | 10.1124 | 1.78326 | 5.63915 | 32.1574 | 1.47054 | 3.16817 | 6.82562 | .314465 |
| 3.19 | 10.1761 | 1.78606 | 5.64801 | 32.4618 | 1.47208 | 3.17149 | 6.83277 | .313480 |
| **3.20** | 10.2400 | 1.78885 | 5.65685 | 32.7680 | 1.47361 | 3.17480 | 6.83990 | .312500 |
| 3.21 | 10.3041 | 1.79165 | 5.66569 | 33.0762 | 1.47515 | 3.17811 | 6.84702 | .311526 |
| 3.22 | 10.3684 | 1.79444 | 5.67450 | 33.3862 | 1.47668 | 3.18140 | 6.85412 | .310559 |
| 3.23 | 10.4329 | 1.79722 | 5.68331 | 33.6983 | 1.47820 | 3.18469 | 6.86121 | .309598 |
| 3.24 | 10.4976 | 1.80000 | 5.69210 | 34.0122 | 1.47973 | 3.18798 | 6.86829 | .308642 |
| 3.25 | 10.5625 | 1.80278 | 5.70088 | 34.3281 | 1.48125 | 3.19125 | 6.87534 | .307692 |
| 3.26 | 10.6276 | 1.80555 | 5.70964 | 34.6460 | 1.48277 | 3.19452 | 6.88239 | .306748 |
| 3.27 | 10.6929 | 1.80831 | 5.71839 | 34.9658 | 1.48428 | 3.19778 | 6.88942 | .305810 |
| 3.28 | 10.7584 | 1.81108 | 5.72713 | 35.2876 | 1.48579 | 3.20104 | 6.89643 | .304878 |
| 3.29 | 10.8241 | 1.81384 | 5.73585 | 35.6113 | 1.48730 | 3.20429 | 6.90344 | .303951 |
| **3.30** | 10.8900 | 1.81659 | 5.74456 | 35.9370 | 1.48881 | 3.20753 | 6.91042 | .303030 |
| 3.31 | 10.9561 | 1.81934 | 5.75326 | 36.2647 | 1.49031 | 3.21077 | 6.91740 | .302115 |
| 3.32 | 11.0224 | 1.82209 | 5.76194 | 36.5944 | 1.49181 | 3.21400 | 6.92436 | .301205 |
| 3.33 | 11.0889 | 1.82483 | 5.77062 | 36.9260 | 1.49330 | 3.21722 | 6.93130 | .300300 |
| 3.34 | 11.1556 | 1.82757 | 5.77927 | 37.2597 | 1.49480 | 3.22044 | 6.93823 | .299401 |
| 3.35 | 11.2225 | 1.83030 | 5.78792 | 37.5954 | 1.49629 | 3.22365 | 6.94515 | .298507 |
| 3.36 | 11.2896 | 1.83303 | 5.79655 | 37.9331 | 1.49777 | 3.22686 | 6.95205 | .297619 |
| 3.37 | 11.3569 | 1.83576 | 5.80517 | 38.2728 | 1.49926 | 3.23006 | 6.95894 | .296736 |
| 3.38 | 11.4244 | 1.83848 | 5.81378 | 38.6145 | 1.50074 | 3.23325 | 6.96582 | .295858 |
| 3.39 | 11.4921 | 1.84120 | 5.82237 | 38.9582 | 1.50222 | 3.23643 | 6.97268 | .294985 |
| **3.40** | 11.5600 | 1.84391 | 5.83095 | 39.3040 | 1.50369 | 3.23961 | 6.97953 | .294118 |
| 3.41 | 11.6281 | 1.84662 | 5.83952 | 39.6518 | 1.50517 | 3.24278 | 6.98637 | .293255 |
| 3.42 | 11.6964 | 1.84932 | 5.84808 | 40.0017 | 1.50664 | 3.24595 | 6.99319 | .292398 |
| 3.43 | 11.7649 | 1.85203 | 5.85662 | 40.3536 | 1.50810 | 3.24911 | 7.00000 | .291545 |
| 3.44 | 11.8336 | 1.85472 | 5.86515 | 40.7076 | 1.50957 | 3.25227 | 7.00680 | .290698 |
| 3.45 | 11.9025 | 1.85742 | 5.87367 | 41.0636 | 1.51103 | 3.25542 | 7.01358 | .289855 |
| 3.46 | 11.9716 | 1.86011 | 5.88218 | 41.4217 | 1.51249 | 3.25856 | 7.02035 | .289017 |
| 3.47 | 12.0409 | 1.86279 | 5.89067 | 41.7819 | 1.51394 | 3.26169 | 7.02711 | .288184 |
| 3.48 | 12.1104 | 1.86548 | 5.89915 | 42.1442 | 1.51540 | 3.26482 | 7.03385 | .287356 |
| 3.49 | 12.1801 | 1.86815 | 5.90762 | 42.5085 | 1.51685 | 3.26795 | 7.04058 | .286533 |
| **3.50** | 12.2500 | 1.87083 | 5.91608 | 42.8750 | 1.51829 | 3.27107 | 7.04730 | .285714 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **3.50** | 12.2500 | 1.87083 | 5.91608 | 42.8750 | 1.51829 | 3.27107 | 7.04730 | .285714 |
| 3.51 | 12.3201 | 1.87350 | 5.92453 | 43.2436 | 1.51974 | 3.27418 | 7.05400 | .284900 |
| 3.52 | 12.3904 | 1.87617 | 5.93296 | 43.6142 | 1.52118 | 3.27729 | 7.06070 | .284091 |
| 3.53 | 12.4609 | 1.87883 | 5.94138 | 43.9870 | 1.52262 | 3.28039 | 7.06738 | .283286 |
| 3.54 | 12.5316 | 1.88149 | 5.94979 | 44.3619 | 1.52406 | 3.28348 | 7.07404 | .282486 |
| 3.55 | 12.6025 | 1.88414 | 5.95819 | 44.7389 | 1.52549 | 3.28657 | 7.08070 | .281690 |
| 3.56 | 12.6736 | 1.88680 | 5.96657 | 45.1180 | 1.52692 | 3.28965 | 7.08734 | .280899 |
| 3.57 | 12.7449 | 1.88944 | 5.97495 | 45.4993 | 1.52835 | 3.29273 | 7.09397 | .280112 |
| 3.58 | 12.8164 | 1.89209 | 5.98331 | 45.8827 | 1.52978 | 3.29580 | 7.10059 | .279330 |
| 3.59 | 12.8881 | 1.89473 | 5.99166 | 46.2683 | 1.53120 | 3.29887 | 7.10719 | .278552 |
| **3.60** | 12.9600 | 1.89737 | 6.00000 | 46.6560 | 1.53262 | 3.30193 | 7.11379 | .277778 |
| 3.61 | 13.0321 | 1.90000 | 6.00833 | 47.0459 | 1.53404 | 3.30498 | 7.12037 | .277008 |
| 3.62 | 13.1044 | 1.90263 | 6.01664 | 47.4379 | 1.53545 | 3.30803 | 7.12694 | .276243 |
| 3.63 | 13.1769 | 1.90526 | 6.02495 | 47.8321 | 1.53686 | 3.31107 | 7.13349 | .275482 |
| 3.64 | 13.2496 | 1.90788 | 6.03324 | 48.2285 | 1.53827 | 3.31411 | 7.14004 | .274725 |
| 3.65 | 13.3225 | 1.91050 | 6.04152 | 48.6271 | 1.53968 | 3.31714 | 7.14657 | .273973 |
| 3.66 | 13.3956 | 1.91311 | 6.04979 | 49.0279 | 1.54109 | 3.32017 | 7.15309 | .273224 |
| 3.67 | 13.4689 | 1.91572 | 6.05805 | 49.4309 | 1.54249 | 3.32319 | 7.15960 | .272480 |
| 3.68 | 13.5424 | 1.91833 | 6.06630 | 49.8360 | 1.54389 | 3.32621 | 7.16610 | .271739 |
| 3.69 | 13.6161 | 1.92094 | 6.07454 | 50.2434 | 1.54529 | 3.32922 | 7.17258 | .271003 |
| **3.70** | 13.6900 | 1.92354 | 6.08276 | 50.6530 | 1.54668 | 3.33222 | 7.17905 | .270270 |
| 3.71 | 13.7641 | 1.92614 | 6.09098 | 51.0648 | 1.54807 | 3.33522 | 7.18552 | .269542 |
| 3.72 | 13.8384 | 1.92873 | 6.09918 | 51.4788 | 1.54946 | 3.33822 | 7.19197 | .268817 |
| 3.73 | 13.9129 | 1.93132 | 6.10737 | 51.8951 | 1.55085 | 3.34120 | 7.19840 | .268097 |
| 3.74 | 13.9876 | 1.93391 | 6.11555 | 52.3136 | 1.55223 | 3.34419 | 7.20483 | .267380 |
| 3.75 | 14.0625 | 1.93649 | 6.12372 | 52.7344 | 1.55362 | 3.34716 | 7.21125 | .266667 |
| 3.76 | 14.1376 | 1.93907 | 6.13188 | 53.1574 | 1.55500 | 3.35014 | 7.21765 | .265957 |
| 3.77 | 14.2129 | 1.94165 | 6.14003 | 53.5826 | 1.55637 | 3.35310 | 7.22405 | .265252 |
| 3.78 | 14.2884 | 1.94422 | 6.14817 | 54.0102 | 1.55775 | 3.35607 | 7.23043 | .264550 |
| 3.79 | 14.3641 | 1.94679 | 6.15630 | 54.4399 | 1.55912 | 3.35902 | 7.23680 | .263852 |
| **3.80** | 14.4400 | 1.94936 | 6.16441 | 54.8720 | 1.56049 | 3.36198 | 7.24316 | .263158 |
| 3.81 | 14.5161 | 1.95192 | 6.17252 | 55.3063 | 1.56186 | 3.36492 | 7.24950 | .262467 |
| 3.82 | 14.5924 | 1.95448 | 6.18061 | 55.7430 | 1.56322 | 3.36786 | 7.25584 | .261780 |
| 3.83 | 14.6689 | 1.95704 | 6.18870 | 56.1819 | 1.56459 | 3.37080 | 7.26217 | .261097 |
| 3.84 | 14.7456 | 1.95959 | 6.19677 | 56.6231 | 1.56595 | 3.37373 | 7.26848 | .260417 |
| 3.85 | 14.8225 | 1.96214 | 6.20484 | 57.0666 | 1.56731 | 3.37666 | 7.27479 | .259740 |
| 3.86 | 14.8996 | 1.96469 | 6.21289 | 57.5125 | 1.56866 | 3.37958 | 7.28108 | .259067 |
| 3.87 | 14.9769 | 1.96723 | 6.22093 | 57.9606 | 1.57001 | 3.38249 | 7.28736 | .258398 |
| 3.88 | 15.0544 | 1.96977 | 6.22896 | 58.4111 | 1.57137 | 3.38540 | 7.29363 | .257732 |
| 3.89 | 15.1321 | 1.97231 | 6.23699 | 58.8639 | 1.57271 | 3.38831 | 7.29989 | .257069 |
| **3.90** | 15.2100 | 1.97484 | 6.24500 | 59.3190 | 1.57406 | 3.39121 | 7.30614 | .256410 |
| 3.91 | 15.2881 | 1.97737 | 6.25300 | 59.7765 | 1.57541 | 3.39411 | 7.31238 | .255754 |
| 3.92 | 15.3664 | 1.97990 | 6.26099 | 60.2363 | 1.57675 | 3.39700 | 7.31861 | .255102 |
| 3.93 | 15.4449 | 1.98242 | 6.26897 | 60.6985 | 1.57809 | 3.39988 | 7.32483 | .254453 |
| 3.94 | 15.5236 | 1.98494 | 6.27694 | 61.1630 | 1.57942 | 3.40277 | 7.33104 | .253807 |
| 3.95 | 15.6025 | 1.98746 | 6.28490 | 61.6299 | 1.58076 | 3.40564 | 7.33723 | .253165 |
| 3.96 | 15.6816 | 1.98997 | 6.29285 | 62.0991 | 1.58209 | 3.40851 | 7.34342 | .252525 |
| 3.97 | 15.7609 | 1.99249 | 6.30079 | 62.5708 | 1.58342 | 3.41138 | 7.34960 | .251889 |
| 3.98 | 15.8404 | 1.99499 | 6.30872 | 63.0448 | 1.58475 | 3.41424 | 7.35576 | .251256 |
| 3.99 | 15.9201 | 1.99750 | 6.31664 | 63.5212 | 1.58608 | 3.41710 | 7.36192 | .250627 |
| **4.00** | 16.0000 | 2.00000 | 6.32456 | 64.0000 | 1.58740 | 3.41995 | 7.36806 | .250000 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **4.00** | 16.0000 | 2.00000 | 6.32456 | 64.0000 | 1.58740 | 3.41995 | 7.36806 | .250000 |
| 4.01 | 16.0801 | 2.00250 | 6.33246 | 64.4812 | 1.58872 | 3.42280 | 7.37420 | .249377 |
| 4.02 | 16.1604 | 2.00499 | 6.34035 | 64.9648 | 1.59004 | 3.42564 | 7.38032 | .248756 |
| 4.03 | 16.2409 | 2.00749 | 6.34823 | 65.4508 | 1.59136 | 3.42848 | 7.38644 | .248139 |
| 4.04 | 16.3216 | 2.00998 | 6.35610 | 65.9393 | 1.59267 | 3.43131 | 7.39254 | .247525 |
| 4.05 | 16.4025 | 2.01246 | 6.36396 | 66.4301 | 1.59399 | 3.43414 | 7.39864 | .246914 |
| 4.06 | 16.4836 | 2.01494 | 6.37181 | 66.9234 | 1.59530 | 3.43697 | 7.40472 | .246305 |
| 4.07 | 16.5649 | 2.01742 | 6.37966 | 67.4191 | 1.59661 | 3.43979 | 7.41080 | .245700 |
| 4.08 | 16.6464 | 2.01990 | 6.38749 | 67.9173 | 1.59791 | 3.44260 | 7.41686 | .245098 |
| 4.09 | 16.7281 | 2.02237 | 6.39531 | 68.4179 | 1.59922 | 3.44541 | 7.42291 | .244499 |
| **4.10** | 16.8100 | 2.02485 | 6.40312 | 68.9210 | 1.60052 | 3.44822 | 7.42896 | .243902 |
| 4.11 | 16.8921 | 2.02731 | 6.41093 | 69.4265 | 1.60182 | 3.45102 | 7.43499 | .243309 |
| 4.12 | 16.9744 | 2.02978 | 6.41872 | 69.9345 | 1.60312 | 3.45382 | 7.44102 | .242718 |
| 4.13 | 17.0569 | 2.03224 | 6.42651 | 70.4450 | 1.60441 | 3.45661 | 7.44703 | .242131 |
| 4.14 | 17.1396 | 2.03470 | 6.43428 | 70.9579 | 1.60571 | 3.45939 | 7.45304 | .241546 |
| 4.15 | 17.2225 | 2.03715 | 6.44205 | 71.4734 | 1.60700 | 3.46218 | 7.45904 | .240964 |
| 4.16 | 17.3056 | 2.03961 | 6.44981 | 71.9913 | 1.60829 | 3.46496 | 7.46502 | .240385 |
| 4.17 | 17.3889 | 2.04206 | 6.45755 | 72.5117 | 1.60958 | 3.46773 | 7.47100 | .239808 |
| 4.18 | 17.4724 | 2.04450 | 6.46529 | 73.0346 | 1.61086 | 3.47050 | 7.47697 | .239234 |
| 4.19 | 17.5561 | 2.04695 | 6.47302 | 73.5601 | 1.61215 | 3.47327 | 7.48292 | .238663 |
| **4.20** | 17.6400 | 2.04939 | 6.48074 | 74.0880 | 1.61343 | 3.47603 | 7.48887 | .238095 |
| 4.21 | 17.7241 | 2.05183 | 6.48845 | 74.6185 | 1.61471 | 3.47878 | 7.49481 | .237530 |
| 4.22 | 17.8084 | 2.05426 | 6.49615 | 75.1514 | 1.61599 | 3.48154 | 7.50074 | .236967 |
| 4.23 | 17.8929 | 2.05670 | 6.50384 | 75.6870 | 1.61726 | 3.48428 | 7.50666 | .236407 |
| 4.24 | 17.9776 | 2.05913 | 6.51153 | 76.2250 | 1.61853 | 3.48703 | 7.51257 | .235849 |
| 4.25 | 18.0625 | 2.06155 | 6.51920 | 76.7656 | 1.61981 | 3.48977 | 7.51847 | .235294 |
| 4.26 | 18.1476 | 2.06398 | 6.52687 | 77.3088 | 1.62108 | 3.49250 | 7.52437 | .234742 |
| 4.27 | 18.2329 | 2.06640 | 6.53452 | 77.8545 | 1.62234 | 3.49523 | 7.53025 | .234192 |
| 4.28 | 18.3184 | 2.06882 | 6.54217 | 78.4028 | 1.62361 | 3.49796 | 7.53612 | .233645 |
| 4.29 | 18.4041 | 2.07123 | 6.54981 | 78.9536 | 1.62487 | 3.50068 | 7.54199 | .233100 |
| **4.30** | 18.4900 | 2.07364 | 6.55744 | 79.5070 | 1.62613 | 3.50340 | 7.54784 | .232558 |
| 4.31 | 18.5761 | 2.07605 | 6.56506 | 80.0630 | 1.62739 | 3.50611 | 7.55369 | .232019 |
| 4.32 | 18.6624 | 2.07846 | 6.57267 | 80.6216 | 1.62865 | 3.50882 | 7.55953 | .231481 |
| 4.33 | 18.7489 | 2.08087 | 6.58027 | 81.1827 | 1.62991 | 3.51153 | 7.56535 | .230947 |
| 4.34 | 18.8356 | 2.08327 | 6.58787 | 81.7465 | 1.63116 | 3.51423 | 7.57117 | .230415 |
| 4.35 | 18.9225 | 2.08567 | 6.59545 | 82.3129 | 1.63241 | 3.51692 | 7.57698 | .229885 |
| 4.36 | 19.0096 | 2.08806 | 6.60303 | 82.8819 | 1.63366 | 3.51962 | 7.58279 | .229358 |
| 4.37 | 19.0969 | 2.09045 | 6.61060 | 83.4535 | 1.63491 | 3.52231 | 7.58858 | .228833 |
| 4.38 | 19.1844 | 2.09284 | 6.61816 | 84.0277 | 1.63619 | 3.52499 | 7.59436 | .228311 |
| 4.39 | 19.2721 | 2.09523 | 6.62571 | 84.6045 | 1.63740 | 3.52767 | 7.60014 | .227790 |
| **4.40** | 19.3600 | 2.09762 | 6.63325 | 85.1840 | 1.63864 | 3.53035 | 7.60590 | .227273 |
| 4.41 | 19.4481 | 2.10000 | 6.64078 | 85.7661 | 1.63988 | 3.53302 | 7.61166 | .226757 |
| 4.42 | 19.5364 | 2.10238 | 6.64831 | 86.3509 | 1.64112 | 3.53569 | 7.61741 | .226244 |
| 4.43 | 19.6249 | 2.10476 | 6.65582 | 86.9383 | 1.64236 | 3.53835 | 7.62315 | .225734 |
| 4.44 | 19.7136 | 2.10713 | 6.66333 | 87.5284 | 1.64359 | 3.54101 | 7.62888 | .225225 |
| 4.45 | 19.8025 | 2.10950 | 6.67083 | 88.1211 | 1.64483 | 3.54367 | 7.63461 | .224719 |
| 4.46 | 19.8916 | 2.11187 | 6.67832 | 88.7165 | 1.64606 | 3.54632 | 7.64032 | .224215 |
| 4.47 | 19.9809 | 2.11424 | 6.68581 | 89.3146 | 1.64729 | 3.54897 | 7.64603 | .223714 |
| 4.48 | 20.0704 | 2.11660 | 6.69328 | 89.9154 | 1.64851 | 3.55162 | 7.65172 | .223214 |
| 4.49 | 20.1601 | 2.11896 | 6.70075 | 90.5188 | 1.64974 | 3.55426 | 7.65741 | .222717 |
| **4.50** | 20.2500 | 2.12132 | 6.70820 | 91.1250 | 1.65096 | 3.55689 | 7.66309 | .222222 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **4.50** | 20.2500 | 2.12132 | 6.70820 | 91.1250 | 1.65096 | 3.55689 | 7.66309 | .222222 |
| 4.51 | 20.3401 | 2.12368 | 6.71565 | 91.7339 | 1.65219 | 3.55953 | 7.66877 | .221729 |
| 4.52 | 20.4304 | 2.12603 | 6.72309 | 92.3454 | 1.65341 | 3.56215 | 7.67443 | .221239 |
| 4.53 | 20.5209 | 2.12838 | 6.73053 | 92.9597 | 1.65462 | 3.56478 | 7.68009 | .220751 |
| 4.54 | 20.6116 | 2.13073 | 6.73795 | 93.5767 | 1.65584 | 3.56740 | 7.68573 | .220264 |
| 4.55 | 20.7025 | 2.13307 | 6.74537 | 94.1964 | 1.65706 | 3.57002 | 7.69137 | .219780 |
| 4.56 | 20.7936 | 2.13542 | 6.75278 | 94.8188 | 1.65827 | 3.57263 | 7.69700 | .219298 |
| 4.57 | 20.8849 | 2.13776 | 6.76018 | 95.4440 | 1.65948 | 3.57524 | 7.70262 | .218818 |
| 4.58 | 20.9764 | 2.14009 | 6.76757 | 96.0719 | 1.66069 | 3.57785 | 7.70824 | .218341 |
| 4.59 | 21.0681 | 2.14243 | 6.77495 | 96.7026 | 1.66190 | 3.58045 | 7.71384 | .217865 |
| **4.60** | 21.1600 | 2.14476 | 6.78233 | 97.3360 | 1.66310 | 3.58305 | 7.71944 | .217391 |
| 4.61 | 21.2521 | 2.14709 | 6.78970 | 97.9722 | 1.66431 | 3.58564 | 7.72503 | .216920 |
| 4.62 | 21.3444 | 2.14942 | 6.79706 | 98.6111 | 1.66551 | 3.58823 | 7.73061 | .216450 |
| 4.63 | 21.4369 | 2.15174 | 6.80441 | 99.2528 | 1.66671 | 3.59082 | 7.73619 | .215983 |
| 4.64 | 21.5296 | 2.15407 | 6.81175 | 99.8673 | 1.66791 | 3.59340 | 7.74175 | .215517 |
| 4.65 | 21.6225 | 2.15639 | 6.81909 | 100.545 | 1.66911 | 3.59598 | 7.74731 | .215054 |
| 4.66 | 21.7156 | 2.15870 | 6.82642 | 101.195 | 1.67030 | 3.59856 | 7.75286 | .214592 |
| 4.67 | 21.8089 | 2.16102 | 6.83374 | 101.848 | 1.67150 | 3.60113 | 7.75840 | .214133 |
| 4.68 | 21.9024 | 2.16333 | 6.84105 | 102.503 | 1.67269 | 3.60370 | 7.76394 | .213675 |
| 4.69 | 21.9961 | 2.16564 | 6.84836 | 103.162 | 1.67388 | 3.60626 | 7.76946 | .213220 |
| **4.70** | 22.0900 | 2.16795 | 6.85565 | 103.823 | 1.67507 | 3.60883 | 7.77498 | .212766 |
| 4.71 | 22.1841 | 2.17025 | 6.86294 | 104.487 | 1.67626 | 3.61138 | 7.78049 | .212314 |
| 4.72 | 22.2784 | 2.17256 | 6.87023 | 105.154 | 1.67744 | 3.61394 | 7.78599 | .211864 |
| 4.73 | 22.3729 | 2.17486 | 6.87750 | 105.824 | 1.67863 | 3.61649 | 7.79149 | .211416 |
| 4.74 | 22.4676 | 2.17715 | 6.88477 | 106.496 | 1.67981 | 3.61903 | 7.79697 | .210970 |
| 4.75 | 22.5625 | 2.17945 | 6.89202 | 107.172 | 1.68099 | 3.62158 | 7.80245 | .210526 |
| 4.76 | 22.6576 | 2.18174 | 6.89928 | 107.850 | 1.68217 | 3.62412 | 7.80793 | .210084 |
| 4.77 | 22.7529 | 2.18403 | 6.90652 | 108.531 | 1.68334 | 3.62665 | 7.81339 | .209644 |
| 4.78 | 22.8484 | 2.18632 | 6.91375 | 109.215 | 1.68452 | 3.62919 | 7.81885 | .209205 |
| 4.79 | 22.9441 | 2.18861 | 6.92098 | 109.902 | 1.68569 | 3.63172 | 7.82429 | .208768 |
| **4.80** | 23.0400 | 2.19089 | 6.92820 | 110.592 | 1.68687 | 3.63424 | 7.82974 | .208333 |
| 4.81 | 23.1361 | 2.19317 | 6.93542 | 111.285 | 1.68804 | 3.63676 | 7.83517 | .207900 |
| 4.82 | 23.2324 | 2.19545 | 6.94262 | 111.980 | 1.68920 | 3.63928 | 7.84059 | .207469 |
| 4.83 | 23.3289 | 2.19773 | 6.94982 | 112.679 | 1.69037 | 3.64180 | 7.84601 | .207039 |
| 4.84 | 23.4256 | 2.20000 | 6.95701 | 113.380 | 1.69154 | 3.64431 | 7.85142 | .206612 |
| 4.85 | 23.5225 | 2.20227 | 6.96419 | 114.084 | 1.69270 | 3.64682 | 7.85683 | .206186 |
| 4.86 | 23.6196 | 2.20454 | 6.97137 | 114.791 | 1.69386 | 3.64932 | 7.86222 | .205761 |
| 4.87 | 23.7169 | 2.20681 | 6.97854 | 115.501 | 1.69503 | 3.65182 | 7.86761 | .205339 |
| 4.88 | 23.8144 | 2.20907 | 6.98570 | 116.214 | 1.69619 | 3.65432 | 7.87299 | .204918 |
| 4.89 | 23.9121 | 2.21133 | 6.99285 | 116.930 | 1.69734 | 3.65681 | 7.87837 | .204499 |
| **4.90** | 24.0100 | 2.21359 | 7.00000 | 117.649 | 1.69850 | 3.65931 | 7.88374 | .204082 |
| 4.91 | 24.1081 | 2.21585 | 7.00714 | 118.371 | 1.69965 | 3.66179 | 7.88909 | .203666 |
| 4.92 | 24.2064 | 2.21811 | 7.01427 | 119.095 | 1.70081 | 3.66428 | 7.89445 | .203252 |
| 4.93 | 24.3049 | 2.22036 | 7.02140 | 119.823 | 1.70196 | 3.66676 | 7.89979 | .202840 |
| 4.94 | 24.4036 | 2.22261 | 7.02851 | 120.554 | 1.70311 | 3.66924 | 7.90513 | .202429 |
| 4.95 | 24.5025 | 2.22486 | 7.03562 | 121.287 | 1.70426 | 3.67171 | 7.91046 | .202020 |
| 4.96 | 24.6016 | 2.22711 | 7.04273 | 122.024 | 1.70540 | 3.67418 | 7.91578 | .201613 |
| 4.97 | 24.7009 | 2.22935 | 7.04982 | 122.763 | 1.70655 | 3.67665 | 7.92110 | .201207 |
| 4.98 | 24.8004 | 2.23159 | 7.05691 | 123.506 | 1.70769 | 3.67911 | 7.92641 | .200803 |
| 4.99 | 24.9001 | 2.23383 | 7.06399 | 124.251 | 1.70884 | 3.68157 | 7.93171 | .200401 |
| **5.00** | 25.0000 | 2.23607 | 7.07107 | 125.000 | 1.70998 | 3.68463 | 7.93701 | .200000 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **5.00** | 25.0000 | 2.23607 | 7.07107 | 125.000 | 1.70998 | 3.68403 | 7.93701 | .200000 |
| 5.01 | 25.1001 | 2.23830 | 7.07814 | 125.752 | 1.71112 | 3.68649 | 7.94229 | .199601 |
| 5.02 | 25.2004 | 2.24054 | 7.08520 | 126.506 | 1.71225 | 3.68894 | 7.94757 | .199203 |
| 5.03 | 25.3009 | 2.24277 | 7.09225 | 127.264 | 1.71339 | 3.69138 | 7.95285 | .198807 |
| 5.04 | 25.4016 | 2.24499 | 7.09930 | 128.024 | 1.71452 | 3.69383 | 7.95811 | .198413 |
| 5.05 | 25.5025 | 2.24722 | 7.10634 | 128.788 | 1.71566 | 3.69627 | 7.96337 | .198020 |
| 5.06 | 25.6036 | 2.24944 | 7.11337 | 129.554 | 1.71679 | 3.69871 | 7.96863 | .197628 |
| 5.07 | 25.7049 | 2.25167 | 7.12039 | 130.324 | 1.71792 | 3.70114 | 7.97387 | .197239 |
| 5.08 | 25.8064 | 2.25389 | 7.12741 | 131.097 | 1.71905 | 3.70357 | 7.97911 | .196850 |
| 5.09 | 25.9081 | 2.25610 | 7.13442 | 131.872 | 1.72017 | 3.70600 | 7.98434 | .196464 |
| **5.10** | 26.0100 | 2.25832 | 7.14143 | 132.651 | 1.72130 | 3.70843 | 7.98957 | .196078 |
| 5.11 | 26.1121 | 2.26053 | 7.14843 | 133.433 | 1.72242 | 3.71085 | 7.99479 | .195695 |
| 5.12 | 26.2144 | 2.26274 | 7.15542 | 134.218 | 1.72355 | 3.71327 | 8.00000 | .195312 |
| 5.13 | 26.3169 | 2.26495 | 7.16240 | 135.006 | 1.72467 | 3.71569 | 8.00520 | .194932 |
| 5.14 | 26.4196 | 2.26716 | 7.16938 | 135.797 | 1.72579 | 3.71810 | 8.01040 | .194553 |
| 5.15 | 26.5225 | 2.26936 | 7.17635 | 136.591 | 1.72691 | 3.72051 | 8.01559 | .194175 |
| 5.16 | 26.6256 | 2.27156 | 7.18331 | 137.388 | 1.72802 | 3.72292 | 8.02078 | .193798 |
| 5.17 | 26.7289 | 2.27376 | 7.19027 | 138.188 | 1.72914 | 3.72532 | 8.02596 | .193424 |
| 5.18 | 26.8324 | 2.27596 | 7.19722 | 138.992 | 1.73025 | 3.72772 | 8.03113 | .193050 |
| 5.19 | 26.9361 | 2.27816 | 7.20417 | 139.798 | 1.73137 | 3.73012 | 8.03629 | .192678 |
| **5.20** | 27.0400 | 2.28035 | 7.21110 | 140.608 | 1.73248 | 3.73251 | 8.04145 | .192308 |
| 5.21 | 27.1441 | 2.28254 | 7.21803 | 141.421 | 1.73359 | 3.73490 | 8.04660 | .191939 |
| 5.22 | 27.2484 | 2.28473 | 7.22496 | 142.237 | 1.73470 | 3.73729 | 8.05175 | .191571 |
| 5.23 | 27.3529 | 2.28692 | 7.23187 | 143.056 | 1.73580 | 3.73968 | 8.05689 | .191205 |
| 5.24 | 27.4576 | 2.28910 | 7.23878 | 143.878 | 1.73691 | 3.74206 | 8.06202 | .190840 |
| 5.25 | 27.5625 | 2.29129 | 7.24569 | 144.703 | 1.73801 | 3.74443 | 8.06714 | .190476 |
| 5.26 | 27.6676 | 2.29347 | 7.25259 | 145.532 | 1.73912 | 3.74681 | 8.07226 | .190114 |
| 5.27 | 27.7729 | 2.29565 | 7.25948 | 146.363 | 1.74022 | 3.74918 | 8.07737 | .189753 |
| 5.28 | 27.8784 | 2.29783 | 7.26636 | 147.198 | 1.74132 | 3.75155 | 8.08248 | .189394 |
| 5.29 | 27.9841 | 2.30000 | 7.27324 | 148.036 | 1.74242 | 3.75392 | 8.08758 | .189036 |
| **5.30** | 28.0900 | 2.30217 | 7.28011 | 148.877 | 1.74351 | 3.75629 | 8.09267 | .188679 |
| 5.31 | 28.1961 | 2.30434 | 7.28697 | 149.721 | 1.74461 | 3.75865 | 8.09776 | .188324 |
| 5.32 | 28.3024 | 2.30651 | 7.29383 | 150.569 | 1.74570 | 3.76101 | 8.10284 | .187970 |
| 5.33 | 28.4089 | 2.30868 | 7.30068 | 151.419 | 1.74680 | 3.76336 | 8.10791 | .187617 |
| 5.34 | 28.5156 | 2.31084 | 7.30753 | 152.273 | 1.74789 | 3.76571 | 8.11298 | .187266 |
| 5.35 | 28.6225 | 2.31301 | 7.31437 | 153.130 | 1.74898 | 3.76806 | 8.11804 | .186916 |
| 5.36 | 28.7296 | 2.31517 | 7.32120 | 153.991 | 1.75007 | 3.77041 | 8.12310 | .186567 |
| 5.37 | 28.8369 | 2.31733 | 7.32803 | 154.854 | 1.75116 | 3.77275 | 8.12814 | .186220 |
| 5.38 | 28.9444 | 2.31948 | 7.33485 | 155.721 | 1.75224 | 3.77509 | 8.13319 | .185874 |
| 5.39 | 29.0521 | 2.32164 | 7.34166 | 156.591 | 1.75333 | 3.77743 | 8.13822 | .185529 |
| **5.40** | 29.1600 | 2.32379 | 7.34847 | 157.464 | 1.75441 | 3.77976 | 8.14325 | .185185 |
| 5.41 | 29.2681 | 2.32594 | 7.35527 | 158.340 | 1.75549 | 3.78209 | 8.14828 | .184843 |
| 5.42 | 29.3764 | 2.32809 | 7.36206 | 159.220 | 1.75657 | 3.78442 | 8.15329 | .184502 |
| 5.43 | 29.4849 | 2.33024 | 7.36885 | 160.103 | 1.75765 | 3.78675 | 8.15831 | .184162 |
| 5.44 | 29.5936 | 2.33238 | 7.37564 | 160.989 | 1.75873 | 3.78907 | 8.16331 | .183824 |
| 5.45 | 29.7025 | 2.33452 | 7.38241 | 161.879 | 1.75981 | 3.79139 | 8.16831 | .183486 |
| 5.46 | 29.8116 | 2.33666 | 7.38918 | 162.771 | 1.76088 | 3.79371 | 8.17330 | .183150 |
| 5.47 | 29.9209 | 2.33880 | 7.39594 | 163.667 | 1.76196 | 3.79603 | 8.17829 | .182815 |
| 5.48 | 30.0304 | 2.34094 | 7.40270 | 164.567 | 1.76303 | 3.79834 | 8.18327 | .182482 |
| 5.49 | 30.1401 | 2.34307 | 7.40945 | 165.469 | 1.76410 | 3.80065 | 8.18824 | .182149 |
| **5.50** | 30.2500 | 2.34521 | 7.41620 | 166.375 | 1.76517 | 3.80295 | 8.19321 | .181818 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **5.50** | 30.2500 | 2.34521 | 7.41620 | 166.375 | 1.76517 | 3.80295 | 8.19321 | .181818 |
| 5.51 | 30.3601 | 2.34734 | 7.42294 | 167.284 | 1.76624 | 3.80526 | 8.19818 | .181488 |
| 5.52 | 30.4704 | 2.34947 | 7.42967 | 168.197 | 1.76731 | 3.80756 | 8.20313 | .181159 |
| 5.53 | 30.5809 | 2.35160 | 7.43640 | 169.112 | 1.76838 | 3.80985 | 8.20808 | .180832 |
| 5.54 | 30.6916 | 2.35372 | 7.44312 | 170.031 | 1.76944 | 3.81215 | 8.21303 | .180505 |
| 5.55 | 30.8025 | 2.35584 | 7.44983 | 170.954 | 1.77051 | 3.81444 | 8.21797 | .180180 |
| 5.56 | 30.9136 | 2.35797 | 7.45654 | 171.880 | 1.77157 | 3.81673 | 8.22290 | .179856 |
| 5.57 | 31.0249 | 2.36008 | 7.46324 | 172.809 | 1.77263 | 3.81902 | 8.22783 | .179533 |
| 5.58 | 31.1364 | 2.36220 | 7.46994 | 173.741 | 1.77369 | 3.82130 | 8.23275 | .179211 |
| 5.59 | 31.2481 | 2.36432 | 7.47663 | 174.677 | 1.77475 | 3.82358 | 8.23766 | .178891 |
| **5.60** | 31.3600 | 2.36643 | 7.48331 | 175.616 | 1.77581 | 3.82586 | 8.24257 | .178571 |
| 5.61 | 31.4721 | 2.36854 | 7.48999 | 176.558 | 1.77686 | 3.82814 | 8.24747 | .178253 |
| 5.62 | 31.5844 | 2.37065 | 7.49667 | 177.504 | 1.77792 | 3.83041 | 8.25237 | .177936 |
| 5.63 | 31.6969 | 2.37276 | 7.50333 | 178.454 | 1.77897 | 3.83268 | 8.25726 | .177620 |
| 5.64 | 31.8096 | 2.37487 | 7.50999 | 179.406 | 1.78003 | 3.83495 | 8.26215 | .177305 |
| 5.65 | 31.9225 | 2.37697 | 7.51665 | 180.362 | 1.78108 | 3.83722 | 8.26703 | .176991 |
| 5.66 | 32.0356 | 2.37908 | 7.52330 | 181.321 | 1.78213 | 3.83948 | 8.27190 | .176678 |
| 5.67 | 32.1489 | 2.38118 | 7.52994 | 182.284 | 1.78318 | 3.84174 | 8.27677 | .176367 |
| 5.68 | 32.2624 | 2.38328 | 7.53658 | 183.250 | 1.78422 | 3.84399 | 8.28164 | .176056 |
| 5.69 | 32.3761 | 2.38537 | 7.54321 | 184.220 | 1.78527 | 3.84625 | 8.28649 | .175747 |
| **5.70** | 32.4900 | 2.38747 | 7.54983 | 185.193 | 1.78632 | 3.84850 | 8.29134 | .175439 |
| 5.71 | 32.6041 | 2.38956 | 7.55645 | 186.169 | 1.78736 | 3.85075 | 8.29619 | .175131 |
| 5.72 | 32.7184 | 2.39165 | 7.56307 | 187.149 | 1.78840 | 3.85300 | 8.30103 | .174825 |
| 5.73 | 32.8329 | 2.39374 | 7.56968 | 188.133 | 1.78944 | 3.85524 | 8.30587 | .174520 |
| 5.74 | 32.9476 | 2.39583 | 7.57628 | 189.119 | 1.79048 | 3.85748 | 8.31069 | .174216 |
| 5.75 | 33.0625 | 2.39792 | 7.58288 | 190.109 | 1.79152 | 3.85972 | 8.31552 | .173913 |
| 5.76 | 33.1776 | 2.40000 | 7.58947 | 191.103 | 1.79256 | 3.86196 | 8.32034 | .173611 |
| 5.77 | 33.2929 | 2.40208 | 7.59605 | 192.100 | 1.79360 | 3.86419 | 8.32515 | .173310 |
| 5.78 | 33.4084 | 2.40416 | 7.60263 | 193.101 | 1.79463 | 3.86642 | 8.32995 | .173010 |
| 5.79 | 33.5241 | 2.40624 | 7.60920 | 194.105 | 1.79567 | 3.86865 | 8.33476 | .172712 |
| **5.80** | 33.6400 | 2.40832 | 7.61577 | 195.112 | 1.79670 | 3.87088 | 8.33955 | .172414 |
| 5.81 | 33.7561 | 2.41039 | 7.62234 | 196.123 | 1.79773 | 3.87310 | 8.34434 | .172117 |
| 5.82 | 33.8724 | 2.41247 | 7.62889 | 197.137 | 1.79876 | 3.87532 | 8.34913 | .171821 |
| 5.83 | 33.9889 | 2.41454 | 7.63544 | 198.155 | 1.79979 | 3.87754 | 8.35390 | .171527 |
| 5.84 | 34.1056 | 2.41661 | 7.64199 | 199.177 | 1.80082 | 3.87975 | 8.35868 | .171233 |
| 5.85 | 34.2225 | 2.41868 | 7.64853 | 200.202 | 1.80185 | 3.88197 | 8.36345 | .170940 |
| 5.86 | 34.3396 | 2.42074 | 7.65506 | 201.230 | 1.80288 | 3.88418 | 8.36821 | .170649 |
| 5.87 | 34.4569 | 2.42281 | 7.66159 | 202.262 | 1.80390 | 3.88639 | 8.37297 | .170358 |
| 5.88 | 34.5744 | 2.42487 | 7.66812 | 203.297 | 1.80492 | 3.88859 | 8.37772 | .170068 |
| 5.89 | 34.6921 | 2.42693 | 7.67463 | 204.336 | 1.80595 | 3.89080 | 8.38247 | .169779 |
| **5.90** | 34.8100 | 2.42899 | 7.68115 | 205.379 | 1.80697 | 3.89300 | 8.38721 | .169492 |
| 5.91 | 34.9281 | 2.43105 | 7.68765 | 206.425 | 1.80799 | 3.89519 | 8.39194 | .169205 |
| 5.92 | 35.0464 | 2.43311 | 7.69415 | 207.475 | 1.80901 | 3.89739 | 8.39667 | .168919 |
| 5.93 | 35.1649 | 2.43516 | 7.70065 | 208.528 | 1.81003 | 3.89958 | 8.40140 | .168634 |
| 5.94 | 35.2836 | 2.43721 | 7.70714 | 209.585 | 1.81104 | 3.90177 | 8.40612 | .168350 |
| 5.95 | 35.4025 | 2.43926 | 7.71362 | 210.645 | 1.81206 | 3.90396 | 8.41083 | .168067 |
| 5.96 | 35.5216 | 2.44131 | 7.72010 | 211.709 | 1.81307 | 3.90615 | 8.41554 | .167785 |
| 5.97 | 35.6409 | 2.44336 | 7.72658 | 212.776 | 1.81409 | 3.90833 | 8.42025 | .167504 |
| 5.98 | 35.7604 | 2.44540 | 7.73305 | 213.847 | 1.81510 | 3.91051 | 8.42494 | .167224 |
| 5.99 | 35.8801 | 2.44745 | 7.73951 | 214.922 | 1.81611 | 3.91269 | 8.42964 | .166945 |
| **6.00** | 36.0000 | 2.44949 | 7.74597 | 216.000 | 1.81712 | 3.91487 | 8.43433 | .166667 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **6.00** | 36.0000 | 2.44949 | 7.74597 | 216.000 | 1.81712 | 3.91487 | 8.43433 | .166667 |
| 6.01 | 36.1201 | 2.45153 | 7.75242 | 217.082 | 1.81813 | 3.91704 | 8.43901 | .166389 |
| 6.02 | 36.2404 | 2.45357 | 7.75887 | 218.167 | 1.81914 | 3.91921 | 8.44369 | .166113 |
| 6.03 | 36.3609 | 2.45561 | 7.76531 | 219.256 | 1.82014 | 3.92138 | 8.44836 | .165837 |
| 6.04 | 36.4816 | 2.45764 | 7.77174 | 220.349 | 1.82115 | 3.92355 | 8.45303 | .165563 |
| 6.05 | 36.6025 | 2.45967 | 7.77817 | 221.445 | 1.82215 | 3.92571 | 8.45769 | .165289 |
| 6.06 | 36.7236 | 2.46171 | 7.78460 | 222.545 | 1.82316 | 3.92787 | 8.46235 | .165017 |
| 6.07 | 36.8449 | 2.46374 | 7.79102 | 223.649 | 1.82416 | 3.93003 | 8.46700 | .164745 |
| 6.08 | 36.9664 | 2.46577 | 7.79744 | 224.756 | 1.82516 | 3.93219 | 8.47165 | .164474 |
| 6.09 | 37.0881 | 2.46779 | 7.80385 | 225.867 | 1.82616 | 3.93434 | 8.47629 | .164204 |
| **6.10** | 37.2100 | 2.46982 | 7.81025 | 226.981 | 1.82716 | 3.93650 | 8.48093 | .163934 |
| 6.11 | 37.3321 | 2.47184 | 7.81665 | 228.099 | 1.82816 | 3.93865 | 8.48556 | .163666 |
| 6.12 | 37.4544 | 2.47386 | 7.82304 | 229.221 | 1.82915 | 3.94079 | 8.49018 | .163399 |
| 6.13 | 37.5769 | 2.47588 | 7.82943 | 230.346 | 1.83015 | 3.94294 | 8.49481 | .163132 |
| 6.14 | 37.6996 | 2.47790 | 7.83582 | 231.476 | 1.83115 | 3.94508 | 8.49942 | .162866 |
| 6.15 | 37.8225 | 2.47992 | 7.84219 | 232.608 | 1.83214 | 3.94722 | 8.50403 | .162602 |
| 6.16 | 37.9456 | 2.48193 | 7.84857 | 233.745 | 1.83313 | 3.94936 | 8.50864 | .162338 |
| 6.17 | 38.0689 | 2.48395 | 7.85493 | 234.885 | 1.83412 | 3.95150 | 8.51324 | .162075 |
| 6.18 | 38.1924 | 2.48596 | 7.86130 | 236.029 | 1.83511 | 3.95363 | 8.51784 | .161812 |
| 6.19 | 38.3161 | 2.48797 | 7.86766 | 237.177 | 1.83610 | 3.95576 | 8.52243 | .161551 |
| **6.20** | 38.4400 | 2.48998 | 7.87401 | 238.328 | 1.83709 | 3.95789 | 8.52702 | .161290 |
| 6.21 | 38.5641 | 2.49199 | 7.88036 | 239.483 | 1.83808 | 3.96002 | 8.53160 | .161031 |
| 6.22 | 38.6884 | 2.49399 | 7.88670 | 240.642 | 1.83906 | 3.96214 | 8.53618 | .160772 |
| 6.23 | 38.8129 | 2.49600 | 7.89303 | 241.804 | 1.84005 | 3.96427 | 8.54075 | .160514 |
| 6.24 | 38.9376 | 2.49800 | 7.89937 | 242.971 | 1.84103 | 3.96638 | 8.54532 | .160256 |
| 6.25 | 39.0625 | 2.50000 | 7.90569 | 244.141 | 1.84202 | 3.96850 | 8.54988 | .160000 |
| 6.26 | 39.1876 | 2.50200 | 7.91202 | 245.314 | 1.84300 | 3.97062 | 8.55444 | .159744 |
| 6.27 | 39.3129 | 2.50400 | 7.91833 | 246.492 | 1.84398 | 3.97273 | 8.55899 | .159490 |
| 6.28 | 39.4384 | 2.50599 | 7.92465 | 247.673 | 1.84496 | 3.97484 | 8.56354 | .159236 |
| 6.29 | 39.5641 | 2.50799 | 7.93095 | 248.858 | 1.84594 | 3.97695 | 8.56808 | .158983 |
| **6.30** | 39.6900 | 2.50998 | 7.93725 | 250.047 | 1.84691 | 3.97906 | 8.57262 | .158730 |
| 6.31 | 39.8161 | 2.51197 | 7.94355 | 251.240 | 1.84789 | 3.98116 | 8.57715 | .158479 |
| 6.32 | 39.9424 | 2.51396 | 7.94984 | 252.436 | 1.84887 | 3.98326 | 8.58168 | .158228 |
| 6.33 | 40.0689 | 2.51595 | 7.95613 | 253.636 | 1.84984 | 3.98536 | 8.58620 | .157978 |
| 6.34 | 40.1956 | 2.51794 | 7.96241 | 254.840 | 1.85082 | 3.98746 | 8.59072 | .157729 |
| 6.35 | 40.3225 | 2.51992 | 7.96869 | 256.048 | 1.85179 | 3.98956 | 8.59524 | .157480 |
| 6.36 | 40.4496 | 2.52190 | 7.97496 | 257.259 | 1.85276 | 3.99165 | 8.59975 | .157233 |
| 6.37 | 40.5769 | 2.52389 | 7.98123 | 258.475 | 1.85373 | 3.99374 | 8.60425 | .156986 |
| 6.38 | 40.7044 | 2.52587 | 7.98749 | 259.694 | 1.85470 | 3.99583 | 8.60875 | .156740 |
| 6.39 | 40.8321 | 2.52784 | 7.99375 | 260.917 | 1.85567 | 3.99792 | 8.61325 | .156495 |
| **6.40** | 40.9600 | 2.52982 | 8.00000 | 262.144 | 1.85664 | 4.00000 | 8.61774 | .156250 |
| 6.41 | 41.0881 | 2.53180 | 8.00625 | 263.375 | 1.85760 | 4.00208 | 8.62222 | .156006 |
| 6.42 | 41.2164 | 2.53377 | 8.01249 | 264.609 | 1.85857 | 4.00416 | 8.62671 | .155763 |
| 6.43 | 41.3449 | 2.53574 | 8.01873 | 265.848 | 1.85953 | 4.00624 | 8.63118 | .155521 |
| 6.44 | 41.4736 | 2.53772 | 8.02496 | 267.090 | 1.86050 | 4.00832 | 8.63566 | .155280 |
| 6.45 | 41.6025 | 2.53969 | 8.03119 | 268.336 | 1.86146 | 4.01039 | 8.64012 | .155039 |
| 6.46 | 41.7316 | 2.54165 | 8.03741 | 269.586 | 1.86242 | 4.01246 | 8.64459 | .154799 |
| 6.47 | 41.8609 | 2.54362 | 8.04363 | 270.840 | 1.86338 | 4.01453 | 8.64904 | .154560 |
| 6.48 | 41.9904 | 2.54558 | 8.04984 | 272.098 | 1.86434 | 4.01660 | 8.65350 | .154321 |
| 6.49 | 42.1201 | 2.54755 | 8.05605 | 273.359 | 1.86530 | 4.01866 | 8.65795 | .154083 |
| **6.50** | 42.2500 | 2.54951 | 8.06226 | 274.625 | 1.86626 | 4.02073 | 8.66239 | .153846 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **6.50** | 42.2500 | 2.54951 | 8.06226 | 274.625 | 1.86626 | 4.02073 | 8.66239 | .153846 |
| 6.51 | 42.3801 | 2.55147 | 8.06846 | 275.894 | 1.86721 | 4.02279 | 8.66683 | .153610 |
| 6.52 | 42.5104 | 2.55343 | 8.07465 | 277.168 | 1.86817 | 4.02485 | 8.67127 | .153374 |
| 6.53 | 42.6409 | 2.55539 | 8.08084 | 278.445 | 1.86912 | 4.02690 | 8.67570 | .153139 |
| 6.54 | 42.7716 | 2.55734 | 8.08703 | 279.726 | 1.87008 | 4.02896 | 8.68012 | .152905 |
| 6.55 | 42.9025 | 2.55930 | 8.09321 | 281.011 | 1.87103 | 4.03101 | 8.68455 | .152672 |
| 6.56 | 43.0336 | 2.56125 | 8.09938 | 282.300 | 1.87198 | 4.03306 | 8.68896 | .152439 |
| 6.57 | 43.1649 | 2.56320 | 8.10555 | 283.593 | 1.87293 | 4.03511 | 8.69338 | .152207 |
| 6.58 | 43.2964 | 2.56515 | 8.11172 | 284.890 | 1.87388 | 4.03715 | 8.69778 | .151976 |
| 6.59 | 43.4281 | 2.56710 | 8.11788 | 286.191 | 1.87483 | 4.03920 | 8.70219 | .151745 |
| **6.60** | 43.5600 | 2.56905 | 8.12404 | 287.496 | 1.87578 | 4.04124 | 8.70659 | .151515 |
| 6.61 | 43.6921 | 2.57099 | 8.13019 | 288.805 | 1.87672 | 4.04328 | 8.71098 | .151286 |
| 6.62 | 43.8244 | 2.57294 | 8.13634 | 290.118 | 1.87767 | 4.04532 | 8.71537 | .151057 |
| 6.63 | 43.9569 | 2.57488 | 8.14248 | 291.434 | 1.87862 | 4.04735 | 8.71976 | .150830 |
| 6.64 | 44.0896 | 2.57682 | 8.14862 | 292.755 | 1.87956 | 4.04939 | 8.72414 | .150602 |
| 6.65 | 44.2225 | 2.57876 | 8.15475 | 294.080 | 1.88050 | 4.05142 | 8.72852 | .150376 |
| 6.66 | 44.3556 | 2.58070 | 8.16088 | 295.408 | 1.88144 | 4.05345 | 8.73289 | .150150 |
| 6.67 | 44.4889 | 2.58263 | 8.16701 | 296.741 | 1.88239 | 4.05548 | 8.73726 | .149925 |
| 6.68 | 44.6224 | 2.58457 | 8.17313 | 298.078 | 1.88333 | 4.05750 | 8.74162 | .149701 |
| 6.69 | 44.7561 | 2.58650 | 8.17924 | 299.418 | 1.88427 | 4.05953 | 8.74598 | .149477 |
| **6.70** | 44.8900 | 2.58844 | 8.18535 | 300.763 | 1.88520 | 4.06155 | 8.75034 | .149254 |
| 6.71 | 45.0241 | 2.59037 | 8.19146 | 302.112 | 1.88614 | 4.06357 | 8.75469 | .149031 |
| 6.72 | 45.1584 | 2.59230 | 8.19756 | 303.464 | 1.88708 | 4.06559 | 8.75904 | .148810 |
| 6.73 | 45.2929 | 2.59422 | 8.20366 | 304.821 | 1.88801 | 4.06760 | 8.76338 | .148588 |
| 6.74 | 45.4276 | 2.59615 | 8.20975 | 306.182 | 1.88895 | 4.06961 | 8.76772 | .148368 |
| 6.75 | 45.5625 | 2.59808 | 8.21584 | 307.547 | 1.88988 | 4.07163 | 8.77205 | .148148 |
| 6.76 | 45.6976 | 2.60000 | 8.22192 | 308.916 | 1.89081 | 4.07364 | 8.77638 | .147929 |
| 6.77 | 45.8329 | 2.60192 | 8.22800 | 310.289 | 1.89175 | 4.07564 | 8.78071 | .147710 |
| 6.78 | 45.9684 | 2.60384 | 8.23408 | 311.666 | 1.89268 | 4.07765 | 8.78503 | .147493 |
| 6.79 | 46.1041 | 2.60576 | 8.24015 | 313.047 | 1.89361 | 4.07965 | 8.78935 | .147275 |
| **6.80** | 46.2400 | 2.60768 | 8.24621 | 314.432 | 1.89454 | 4.08166 | 8.79366 | .147059 |
| 6.81 | 46.3761 | 2.60960 | 8.25227 | 315.821 | 1.89546 | 4.08365 | 8.79797 | .146843 |
| 6.82 | 46.5124 | 2.61151 | 8.25833 | 317.215 | 1.89639 | 4.08565 | 8.80227 | .146628 |
| 6.83 | 46.6489 | 2.61343 | 8.26438 | 318.612 | 1.89732 | 4.08765 | 8.80657 | .146413 |
| 6.84 | 46.7856 | 2.61534 | 8.27043 | 320.014 | 1.89824 | 4.08964 | 8.81087 | .146199 |
| 6.85 | 46.9225 | 2.61725 | 8.27647 | 321.419 | 1.89917 | 4.09163 | 8.81516 | .145985 |
| 6.86 | 47.0596 | 2.61916 | 8.28251 | 322.829 | 1.90009 | 4.09362 | 8.81945 | .145773 |
| 6.87 | 47.1969 | 2.62107 | 8.28855 | 324.243 | 1.90102 | 4.09561 | 8.82373 | .145560 |
| 6.88 | 47.3344 | 2.62298 | 8.29458 | 325.661 | 1.90194 | 4.09760 | 8.82801 | .145349 |
| 6.89 | 47.4721 | 2.62488 | 8.30060 | 327.083 | 1.90286 | 4.09958 | 8.83228 | .145138 |
| **6.90** | 47.6100 | 2.62679 | 8.30662 | 328.509 | 1.90378 | 4.10157 | 8.83656 | .144928 |
| 6.91 | 47.7481 | 2.62869 | 8.31264 | 329.939 | 1.90470 | 4.10355 | 8.84082 | .144718 |
| 6.92 | 47.8864 | 2.63059 | 8.31865 | 331.374 | 1.90562 | 4.10552 | 8.84509 | .144509 |
| 6.93 | 48.0249 | 2.63249 | 8.32466 | 332.813 | 1.90653 | 4.10750 | 8.84934 | .144300 |
| 6.94 | 48.1636 | 2.63439 | 8.33067 | 334.255 | 1.90745 | 4.10948 | 8.85360 | .144092 |
| 6.95 | 48.3025 | 2.63629 | 8.33667 | 335.702 | 1.90837 | 4.11145 | 8.85785 | .143885 |
| 6.96 | 48.4416 | 2.63818 | 8.34266 | 337.154 | 1.90928 | 4.11342 | 8.86210 | .143678 |
| 6.97 | 48.5809 | 2.64008 | 8.34865 | 338.609 | 1.91019 | 4.11539 | 8.86634 | .143472 |
| 6.98 | 48.7204 | 2.64197 | 8.35464 | 340.068 | 1.91111 | 4.11736 | 8.87058 | .143266 |
| 6.99 | 48.8601 | 2.64386 | 8.36062 | 341.532 | 1.91202 | 4.11932 | 8.87481 | .143062 |
| **7.00** | 49.0000 | 2.64575 | 8.36660 | 343.000 | 1.91293 | 4.12129 | 8.87904 | .142857 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **7.00** | 49.0000 | 2.64575 | 8.36660 | 343.000 | 1.91293 | 4.12129 | 8.87904 | .142857 |
| 7.01 | 49.1401 | 2.64764 | 8.37257 | 344.472 | 1.91384 | 4.12325 | 8.88327 | .142653 |
| 7.02 | 49.2804 | 2.64953 | 8.37854 | 345.948 | 1.91475 | 4.12521 | 8.88749 | .142450 |
| 7.03 | 49.4209 | 2.65141 | 8.38451 | 347.429 | 1.91566 | 4.12716 | 8.89171 | .142248 |
| 7.04 | 49.5616 | 2.65330 | 8.39047 | 348.914 | 1.91657 | 4.12912 | 8.89592 | .142045 |
| 7.05 | 49.7025 | 2.65518 | 8.39643 | 350.403 | 1.91747 | 4.13107 | 8.90013 | .141844 |
| 7.06 | 49.8436 | 2.65707 | 8.40238 | 351.896 | 1.91838 | 4.13303 | 8.90434 | .141643 |
| 7.07 | 49.9849 | 2.65895 | 8.40833 | 353.393 | 1.91929 | 4.13498 | 8.90854 | .141443 |
| 7.08 | 50.1264 | 2.66083 | 8.41427 | 354.895 | 1.92019 | 4.13693 | 8.91274 | .141243 |
| 7.09 | 50.2681 | 2.66271 | 8.42021 | 356.401 | 1.92109 | 4.13887 | 8.91693 | .141044 |
| **7.10** | 50.4100 | 2.66458 | 8.42615 | 357.911 | 1.92200 | 4.14082 | 8.92112 | .140845 |
| 7.11 | 50.5521 | 2.66646 | 8.43208 | 359.425 | 1.92290 | 4.14276 | 8.92531 | .140647 |
| 7.12 | 50.6944 | 2.66833 | 8.43801 | 360.944 | 1.92380 | 4.14470 | 8.92949 | .140449 |
| 7.13 | 50.8369 | 2.67021 | 8.44393 | 362.467 | 1.92470 | 4.14664 | 8.93367 | .140252 |
| 7.14 | 50.9796 | 2.67208 | 8.44985 | 363.994 | 1.92560 | 4.14858 | 8.93784 | .140056 |
| 7.15 | 51.1225 | 2.67395 | 8.45577 | 365.526 | 1.92650 | 4.15052 | 8.94201 | .139860 |
| 7.16 | 51.2656 | 2.67582 | 8.46168 | 367.062 | 1.92740 | 4.15245 | 8.94618 | .139665 |
| 7.17 | 51.4089 | 2.67769 | 8.46759 | 368.602 | 1.92829 | 4.15438 | 8.95034 | .139470 |
| 7.18 | 51.5524 | 2.67955 | 8.47349 | 370.146 | 1.92919 | 4.15631 | 8.95450 | .139276 |
| 7.19 | 51.6961 | 2.68142 | 8.47939 | 371.695 | 1.93008 | 4.15824 | 8.95866 | .139082 |
| **7.20** | 51.8400 | 2.68328 | 8.48528 | 373.248 | 1.93098 | 4.16017 | 8.96281 | .138889 |
| 7.21 | 51.9841 | 2.68514 | 8.49117 | 374.805 | 1.93187 | 4.16209 | 8.96696 | .138696 |
| 7.22 | 52.1284 | 2.68701 | 8.49706 | 376.367 | 1.93277 | 4.16402 | 8.97110 | .138504 |
| 7.23 | 52.2729 | 2.68887 | 8.50294 | 377.933 | 1.93366 | 4.16594 | 8.97524 | .138313 |
| 7.24 | 52.4176 | 2.69072 | 8.50882 | 379.503 | 1.93455 | 4.16786 | 8.97938 | .138122 |
| 7.25 | 52.5625 | 2.69258 | 8.51469 | 381.078 | 1.93544 | 4.16978 | 8.98351 | .137931 |
| 7.26 | 52.7076 | 2.69444 | 8.52056 | 382.657 | 1.93633 | 4.17169 | 8.98764 | .137741 |
| 7.27 | 52.8529 | 2.69629 | 8.52643 | 384.241 | 1.93722 | 4.17361 | 8.99176 | .137552 |
| 7.28 | 52.9984 | 2.69815 | 8.53229 | 385.828 | 1.93810 | 4.17552 | 8.99588 | .137363 |
| 7.29 | 53.1441 | 2.70000 | 8.53815 | 387.420 | 1.93899 | 4.17743 | 9.00000 | .137174 |
| **7.30** | 53.2900 | 2.70185 | 8.54400 | 389.017 | 1.93988 | 4.17934 | 9.00411 | .136986 |
| 7.31 | 53.4361 | 2.70370 | 8.54985 | 390.618 | 1.94076 | 4.18125 | 9.00822 | .136799 |
| 7.32 | 53.5824 | 2.70555 | 8.55570 | 392.223 | 1.94165 | 4.18315 | 9.01233 | .136612 |
| 7.33 | 53.7289 | 2.70740 | 8.56154 | 393.833 | 1.94253 | 4.18506 | 9.01643 | .136426 |
| 7.34 | 53.8756 | 2.70924 | 8.56738 | 395.447 | 1.94341 | 4.18696 | 9.02053 | .136240 |
| 7.35 | 54.0225 | 2.71109 | 8.57321 | 397.065 | 1.94430 | 4.18886 | 9.02462 | .136054 |
| 7.36 | 54.1696 | 2.71293 | 8.57904 | 398.688 | 1.94518 | 4.19076 | 9.02871 | .135870 |
| 7.37 | 54.3169 | 2.71477 | 8.58487 | 400.316 | 1.94606 | 4.19266 | 9.03280 | .135685 |
| 7.38 | 54.4644 | 2.71662 | 8.59069 | 401.947 | 1.94694 | 4.19455 | 9.03689 | .135501 |
| 7.39 | 54.6121 | 2.71846 | 8.59651 | 403.583 | 1.94782 | 4.19644 | 9.04097 | .135318 |
| **7.40** | 54.7600 | 2.72029 | 8.60233 | 405.224 | 1.94870 | 4.19834 | 9.04504 | .135135 |
| 7.41 | 54.9081 | 2.72213 | 8.60814 | 406.869 | 1.94957 | 4.20023 | 9.04911 | .134953 |
| 7.42 | 55.0564 | 2.72397 | 8.61394 | 408.518 | 1.95045 | 4.20212 | 9.05318 | .134771 |
| 7.43 | 55.2049 | 2.72580 | 8.61974 | 410.172 | 1.95132 | 4.20400 | 9.05725 | .134590 |
| 7.44 | 55.3536 | 2.72764 | 8.62554 | 411.831 | 1.95220 | 4.20589 | 9.06131 | .134409 |
| 7.45 | 55.5025 | 2.72947 | 8.63134 | 413.494 | 1.95307 | 4.20777 | 9.06537 | .134228 |
| 7.46 | 55.6516 | 2.73130 | 8.63713 | 415.161 | 1.95395 | 4.20965 | 9.06942 | .134048 |
| 7.47 | 55.8009 | 2.73313 | 8.64292 | 416.833 | 1.95482 | 4.21153 | 9.07347 | .133869 |
| 7.48 | 55.9504 | 2.73496 | 8.64870 | 418.509 | 1.95569 | 4.21341 | 9.07752 | .133690 |
| 7.49 | 56.1001 | 2.73679 | 8.65448 | 420.190 | 1.95656 | 4.21529 | 9.08156 | .133511 |
| **7.50** | 56.2500 | 2.73861 | 8.66025 | 421.875 | 1.95743 | 4.21716 | 9.08560 | .133333 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **7.50** | 56.2500 | 2.73861 | 8.66025 | 421.875 | 1.95743 | 4.21716 | 9.08560 | .133333 |
| 7.51 | 56.4001 | 2.74044 | 8.66603 | 423.565 | 1.95830 | 4.21904 | 9.08964 | .133156 |
| 7.52 | 56.5504 | 2.74226 | 8.67179 | 425.259 | 1.95917 | 4.22091 | 9.09367 | .132979 |
| 7.53 | 56.7009 | 2.74408 | 8.67756 | 426.958 | 1.96004 | 4.22278 | 9.09770 | .132802 |
| 7.54 | 56.8516 | 2.74591 | 8.68332 | 428.661 | 1.96091 | 4.22465 | 9.10173 | .132626 |
| 7.55 | 57.0025 | 2.74773 | 8.68907 | 430.369 | 1.96177 | 4.22651 | 9.10575 | .132450 |
| 7.56 | 57.1536 | 2.74955 | 8.69483 | 432.081 | 1.96264 | 4.22838 | 9.10977 | .132275 |
| 7.57 | 57.3049 | 2.75136 | 8.70057 | 433.798 | 1.96350 | 4.23024 | 9.11378 | .132100 |
| 7.58 | 57.4564 | 2.75318 | 8.70632 | 435.520 | 1.96437 | 4.23210 | 9.11779 | .131926 |
| 7.59 | 57.6081 | 2.75500 | 8.71206 | 437.245 | 1.96523 | 4.23396 | 9.12180 | .131752 |
| **7.60** | 57.7600 | 2.75681 | 8.71780 | 438.976 | 1.96610 | 4.23582 | 9.12581 | .131579 |
| 7.61 | 57.9121 | 2.75862 | 8.72353 | 440.711 | 1.96696 | 4.23768 | 9.12981 | .131406 |
| 7.62 | 58.0644 | 2.76043 | 8.72926 | 442.451 | 1.96782 | 4.23954 | 9.13380 | .131234 |
| 7.63 | 58.2169 | 2.76225 | 8.73499 | 444.195 | 1.96868 | 4.24139 | 9.13780 | .131062 |
| 7.64 | 58.3696 | 2.76405 | 8.74071 | 445.944 | 1.96954 | 4.24324 | 9.14179 | .130890 |
| 7.65 | 58.5225 | 2.76586 | 8.74643 | 447.697 | 1.97040 | 4.24509 | 9.14577 | .130719 |
| 7.66 | 58.6756 | 2.76767 | 8.75214 | 449.455 | 1.97126 | 4.24694 | 9.14976 | .130548 |
| 7.67 | 58.8289 | 2.76948 | 8.75785 | 451.218 | 1.97211 | 4.24879 | 9.15374 | .130378 |
| 7.68 | 58.9824 | 2.77128 | 8.76356 | 452.985 | 1.97297 | 4.25063 | 9.15771 | .130208 |
| 7.69 | 59.1361 | 2.77308 | 8.76926 | 454.757 | 1.97383 | 4.25248 | 9.16169 | .130039 |
| **7.70** | 59.2900 | 2.77489 | 8.77496 | 456.533 | 1.97468 | 4.25432 | 9.16566 | .129870 |
| 7.71 | 59.4441 | 2.77669 | 8.78066 | 458.314 | 1.97554 | 4.25616 | 9.16962 | .129702 |
| 7.72 | 59.5984 | 2.77849 | 8.78635 | 460.100 | 1.97639 | 4.25800 | 9.17359 | .129534 |
| 7.73 | 59.7529 | 2.78029 | 8.79204 | 461.890 | 1.97724 | 4.25984 | 9.17754 | .129366 |
| 7.74 | 59.9076 | 2.78209 | 8.79773 | 463.685 | 1.97809 | 4.26167 | 9.18150 | .129199 |
| 7.75 | 60.0625 | 2.78388 | 8.80341 | 465.484 | 1.97895 | 4.26351 | 9.18545 | .129032 |
| 7.76 | 60.2176 | 2.78568 | 8.80909 | 467.289 | 1.97980 | 4.26534 | 9.18940 | .128866 |
| 7.77 | 60.3729 | 2.78747 | 8.81476 | 469.097 | 1.98065 | 4.26717 | 9.19335 | .128700 |
| 7.78 | 60.5284 | 2.78927 | 8.82043 | 470.911 | 1.98150 | 4.26900 | 9.19729 | .128535 |
| 7.79 | 60.6841 | 2.79106 | 8.82610 | 472.729 | 1.98234 | 4.27083 | 9.20123 | .128370 |
| **7.80** | 60.8400 | 2.79285 | 8.83176 | 474.552 | 1.98319 | 4.27266 | 9.20516 | .128205 |
| 7.81 | 60.9961 | 2.79464 | 8.83742 | 476.380 | 1.98404 | 4.27448 | 9.20910 | .128041 |
| 7.82 | 61.1524 | 2.79643 | 8.84308 | 478.212 | 1.98489 | 4.27631 | 9.21302 | .127877 |
| 7.83 | 61.3089 | 2.79821 | 8.84873 | 480.049 | 1.98573 | 4.27813 | 9.21695 | .127714 |
| 7.84 | 61.4656 | 2.80000 | 8.85438 | 481.890 | 1.98658 | 4.27995 | 9.22087 | .127551 |
| 7.85 | 61.6225 | 2.80179 | 8.86002 | 483.737 | 1.98742 | 4.28177 | 9.22479 | .127389 |
| 7.86 | 61.7796 | 2.80357 | 8.86566 | 485.588 | 1.98826 | 4.28359 | 9.22871 | .127226 |
| 7.87 | 61.9369 | 2.80535 | 8.87130 | 487.443 | 1.98911 | 4.28540 | 9.23262 | .127065 |
| 7.88 | 62.0944 | 2.80713 | 8.87694 | 489.304 | 1.98995 | 4.28722 | 9.23653 | .126904 |
| 7.89 | 62.2521 | 2.80891 | 8.88257 | 491.169 | 1.99079 | 4.28903 | 9.24043 | .126743 |
| **7.90** | 62.4100 | 2.81069 | 8.88819 | 493.039 | 1.99163 | 4.29084 | 9.24434 | .126582 |
| 7.91 | 62.5681 | 2.81247 | 8.89382 | 494.914 | 1.99247 | 4.29265 | 9.24823 | .126422 |
| 7.92 | 62.7264 | 2.81425 | 8.89944 | 496.793 | 1.99331 | 4.29446 | 9.25213 | .126263 |
| 7.93 | 62.8849 | 2.81603 | 8.90505 | 498.677 | 1.99415 | 4.29627 | 9.25602 | .126103 |
| 7.94 | 63.0436 | 2.81780 | 8.91067 | 500.566 | 1.99499 | 4.29807 | 9.25991 | .125945 |
| 7.95 | 63.2025 | 2.81957 | 8.91628 | 502.460 | 1.99582 | 4.29987 | 9.26380 | .125786 |
| 7.96 | 63.3616 | 2.82135 | 8.92188 | 504.358 | 1.99666 | 4.30168 | 9.26768 | .125628 |
| 7.97 | 63.5209 | 2.82312 | 8.92749 | 506.262 | 1.99750 | 4.30348 | 9.27156 | .125471 |
| 7.98 | 63.6804 | 2.82489 | 8.93308 | 508.170 | 1.99833 | 4.30528 | 9.27544 | .125313 |
| 7.99 | 63.8401 | 2.82666 | 8.93868 | 510.082 | 1.99917 | 4.30707 | 9.27931 | .125156 |
| **8.00** | 64.0000 | 2.82843 | 8.94427 | 512.000 | 2.00000 | 4.30887 | 9.28318 | .125000 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **8.00** | 64.0000 | 2.82843 | 8.94427 | 512.000 | 2.00000 | 4.30887 | 9.28318 | .125000 |
| 8.01 | 64.1601 | 2.83019 | 8.94986 | 513.922 | 2.00083 | 4.31066 | 9.28704 | .124844 |
| 8.02 | 64.3204 | 2.83196 | 8.95545 | 515.850 | 2.00167 | 4.31246 | 9.29091 | .124688 |
| 8.03 | 64.4809 | 2.83373 | 8.96103 | 517.782 | 2.00250 | 4.31425 | 9.29477 | .124533 |
| 8.04 | 64.6416 | 2.83549 | 8.96660 | 519.718 | 2.00333 | 4.31604 | 9.29862 | .124378 |
| 8.05 | 64.8025 | 2.83725 | 8.97218 | 521.660 | 2.00416 | 4.31783 | 9.30248 | .124224 |
| 8.06 | 64.9636 | 2.83901 | 8.97775 | 523.607 | 2.00499 | 4.31961 | 9.30633 | .124069 |
| 8.07 | 65.1249 | 2.84077 | 8.98332 | 525.558 | 2.00582 | 4.32140 | 9.31018 | .123916 |
| 8.08 | 65.2864 | 2.84253 | 8.98888 | 527.514 | 2.00664 | 4.32318 | 9.31402 | .123762 |
| 8.09 | 65.4481 | 2.84429 | 8.99444 | 529.475 | 2.00747 | 4.32497 | 9.31786 | .123609 |
| **8.10** | 65.6100 | 2.84605 | 9.00000 | 531.441 | 2.00830 | 4.32675 | 9.32170 | .123457 |
| 8.11 | 65.7721 | 2.84781 | 9.00555 | 533.412 | 2.00912 | 4.32853 | 9.32553 | .123305 |
| 8.12 | 65.9344 | 2.84956 | 9.01110 | 535.387 | 2.00995 | 4.33031 | 9.32936 | .123153 |
| 8.13 | 66.0969 | 2.85132 | 9.01665 | 537.368 | 2.01078 | 4.33208 | 9.33319 | .123001 |
| 8.14 | 66.2596 | 2.85307 | 9.02219 | 539.353 | 2.01160 | 4.33386 | 9.33702 | .122850 |
| 8.15 | 66.4225 | 2.85482 | 9.02774 | 541.343 | 2.01242 | 4.33563 | 9.34084 | .122699 |
| 8.16 | 66.5856 | 2.85657 | 9.03327 | 543.338 | 2.01325 | 4.33741 | 9.34466 | .122549 |
| 8.17 | 66.7489 | 2.85832 | 9.03881 | 545.339 | 2.01407 | 4.33918 | 9.34847 | .122399 |
| 8.18 | 66.9124 | 2.86007 | 9.04434 | 547.343 | 2.01489 | 4.34095 | 9.35229 | .122249 |
| 8.19 | 67.0761 | 2.86182 | 9.04986 | 549.353 | 2.01571 | 4.34271 | 9.35610 | .122100 |
| **8.20** | 67.2400 | 2.86356 | 9.05539 | 551.368 | 2.01653 | 4.34448 | 9.35990 | .121951 |
| 8.21 | 67.4041 | 2.86531 | 9.06091 | 553.388 | 2.01735 | 4.34625 | 9.36370 | .121803 |
| 8.22 | 67.5684 | 2.86705 | 9.06642 | 555.412 | 2.01817 | 4.34801 | 9.36751 | .121655 |
| 8.23 | 67.7329 | 2.86880 | 9.07193 | 557.442 | 2.01899 | 4.34977 | 9.37130 | .121507 |
| 8.24 | 67.8976 | 2.87054 | 9.07744 | 559.476 | 2.01980 | 4.35153 | 9.37510 | .121359 |
| 8.25 | 68.0625 | 2.87228 | 9.08295 | 561.516 | 2.02062 | 4.35329 | 9.37889 | .121212 |
| 8.26 | 68.2276 | 2.87402 | 9.08845 | 563.560 | 2.02144 | 4.35505 | 9.38268 | .121065 |
| 8.27 | 68.3929 | 2.87576 | 9.09395 | 565.609 | 2.02225 | 4.35681 | 9.38646 | .120919 |
| 8.28 | 68.5584 | 2.87750 | 9.09945 | 567.664 | 2.02307 | 4.35856 | 9.39024 | .120773 |
| 8.29 | 68.7241 | 2.87924 | 9.10494 | 569.723 | 2.02388 | 4.36032 | 9.39402 | .120627 |
| **8.30** | 68.8900 | 2.88097 | 9.11043 | 571.787 | 2.02469 | 4.36207 | 9.39780 | .120482 |
| 8.31 | 69.0561 | 2.88271 | 9.11592 | 573.856 | 2.02551 | 4.36382 | 9.40157 | .120337 |
| 8.32 | 69.2224 | 2.88444 | 9.12140 | 575.930 | 2.02632 | 4.36557 | 9.40534 | .120192 |
| 8.33 | 69.3889 | 2.88617 | 9.12688 | 578.010 | 2.02713 | 4.36732 | 9.40911 | .120048 |
| 8.34 | 69.5556 | 2.88791 | 9.13236 | 580.094 | 2.02794 | 4.36907 | 9.41287 | .119904 |
| 8.35 | 69.7225 | 2.88964 | 9.13783 | 582.183 | 2.02875 | 4.37081 | 9.41663 | .119760 |
| 8.36 | 69.8896 | 2.89137 | 9.14330 | 584.277 | 2.02956 | 4.37256 | 9.42039 | .119617 |
| 8.37 | 70.0569 | 2.89310 | 9.14877 | 586.376 | 2.03037 | 4.37430 | 9.42414 | .119474 |
| 8.38 | 70.2244 | 2.89482 | 9.15423 | 588.480 | 2.03118 | 4.37604 | 9.42789 | .119332 |
| 8.39 | 70.3921 | 2.89655 | 9.15969 | 590.590 | 2.03199 | 4.37778 | 9.43164 | .119190 |
| **8.40** | 70.5600 | 2.89828 | 9.16515 | 592.704 | 2.03279 | 4.37952 | 9.43539 | .119048 |
| 8.41 | 70.7281 | 2.90000 | 9.17061 | 594.823 | 2.03360 | 4.38126 | 9.43913 | .118906 |
| 8.42 | 70.8964 | 2.90172 | 9.17606 | 596.948 | 2.03440 | 4.38299 | 9.44287 | .118765 |
| 8.43 | 71.0649 | 2.90345 | 9.18150 | 599.077 | 2.03521 | 4.38473 | 9.44661 | .118624 |
| 8.44 | 71.2336 | 2.90517 | 9.18695 | 601.212 | 2.03601 | 4.38646 | 9.45034 | .118483 |
| 8.45 | 71.4025 | 2.90689 | 9.19239 | 603.351 | 2.03682 | 4.38819 | 9.45407 | .118343 |
| 8.46 | 71.5716 | 2.90861 | 9.19783 | 605.496 | 2.03762 | 4.38992 | 9.45780 | .118203 |
| 8.47 | 71.7409 | 2.91033 | 9.20326 | 607.645 | 2.03842 | 4.39165 | 9.46152 | .118064 |
| 8.48 | 71.9104 | 2.91204 | 9.20869 | 609.800 | 2.03923 | 4.39338 | 9.46525 | .117925 |
| 8.49 | 72.0801 | 2.91376 | 9.21412 | 611.960 | 2.04003 | 4.39510 | 9.46897 | .117786 |
| **8.50** | 72.2500 | 2.91548 | 9.21954 | 614.125 | 2.04083 | 4.39683 | 9.47268 | .117647 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **8.50** | 72.2500 | 2.91548 | 9.21954 | 614.125 | 2.04083 | 4.39683 | 9.47268 | .117647 |
| 8.51 | 72.4201 | 2.91719 | 9.22497 | 616.295 | 2.04163 | 4.39855 | 9.47640 | .117509 |
| 8.52 | 72.5904 | 2.91890 | 9.23038 | 618.470 | 2.04243 | 4.40028 | 9.48011 | .117371 |
| 8.53 | 72.7609 | 2.92062 | 9.23580 | 620.650 | 2.04323 | 4.40200 | 9.48381 | .117233 |
| 8.54 | 72.9316 | 2.92233 | 9.24121 | 622.836 | 2.04402 | 4.40372 | 9.48752 | .117096 |
| 8.55 | 73.1025 | 2.92404 | 9.24662 | 625.026 | 2.04482 | 4.40543 | 9.49122 | .116959 |
| 8.56 | 73.2736 | 2.92575 | 9.25203 | 627.222 | 2.04562 | 4.40715 | 9.49492 | .116822 |
| 8.57 | 73.4449 | 2.92746 | 9.25743 | 629.423 | 2.04641 | 4.40887 | 9.49861 | .116686 |
| 8.58 | 73.6164 | 2.92916 | 9.26283 | 631.629 | 2.04721 | 4.41058 | 9.50231 | .116550 |
| 8.59 | 73.7881 | 2.93087 | 9.26823 | 633.840 | 2.04801 | 4.41229 | 9.50600 | .116414 |
| **8.60** | 73.9600 | 2.93258 | 9.27362 | 636.056 | 2.04880 | 4.41400 | 9.50969 | .116279 |
| 8.61 | 74.1321 | 2.93428 | 9.27901 | 638.277 | 2.04959 | 4.41571 | 9.51337 | .116144 |
| 8.62 | 74.3044 | 2.93598 | 9.28440 | 640.504 | 2.05039 | 4.41742 | 9.51705 | .116009 |
| 8.63 | 74.4769 | 2.93769 | 9.28978 | 642.736 | 2.05118 | 4.41913 | 9.52073 | .115875 |
| 8.64 | 74.6496 | 2.93939 | 9.29516 | 644.973 | 2.05197 | 4.42084 | 9.52441 | .115741 |
| 8.65 | 74.8225 | 2.94109 | 9.30054 | 647.215 | 2.05276 | 4.42254 | 9.52808 | .115607 |
| 8.66 | 74.9956 | 2.94279 | 9.30591 | 649.462 | 2.05355 | 4.42425 | 9.53175 | .115473 |
| 8.67 | 75.1689 | 2.94449 | 9.31128 | 651.714 | 2.05434 | 4.42595 | 9.53542 | .115340 |
| 8.68 | 75.3424 | 2.94618 | 9.31665 | 653.972 | 2.05513 | 4.42765 | 9.53908 | .115207 |
| 8.69 | 75.5161 | 2.94788 | 9.32202 | 656.235 | 2.05592 | 4.42935 | 9.54274 | .115075 |
| **8.70** | 75.6900 | 2.94958 | 9.32738 | 658.503 | 2.05671 | 4.43105 | 9.54640 | .114943 |
| 8.71 | 75.8641 | 2.95127 | 9.33274 | 660.776 | 2.05750 | 4.43274 | 9.55006 | .114811 |
| 8.72 | 76.0384 | 2.95296 | 9.33809 | 663.055 | 2.05828 | 4.43444 | 9.55371 | .114679 |
| 8.73 | 76.2129 | 2.95466 | 9.34345 | 665.339 | 2.05907 | 4.43613 | 9.55736 | .114548 |
| 8.74 | 76.3876 | 2.95635 | 9.34880 | 667.628 | 2.05986 | 4.43783 | 9.56101 | .114416 |
| 8.75 | 76.5625 | 2.95804 | 9.35414 | 669.922 | 2.06064 | 4.43952 | 9.56466 | .114286 |
| 8.76 | 76.7376 | 2.95973 | 9.35949 | 672.221 | 2.06143 | 4.44121 | 9.56830 | .114155 |
| 8.77 | 76.9129 | 2.96142 | 9.36483 | 674.526 | 2.06221 | 4.44290 | 9.57194 | .114025 |
| 8.78 | 77.0884 | 2.96311 | 9.37017 | 676.836 | 2.06299 | 4.44459 | 9.57557 | .113895 |
| 8.79 | 77.2641 | 2.96479 | 9.37550 | 679.151 | 2.06378 | 4.44627 | 9.57921 | .113766 |
| **8.80** | 77.4400 | 2.96648 | 9.38083 | 681.472 | 2.06456 | 4.44796 | 9.58284 | .113636 |
| 8.81 | 77.6161 | 2.96816 | 9.38616 | 683.798 | 2.06534 | 4.44964 | 9.58647 | .113507 |
| 8.82 | 77.7924 | 2.96985 | 9.39149 | 686.129 | 2.06612 | 4.45133 | 9.59009 | .113379 |
| 8.83 | 77.9689 | 2.97153 | 9.39681 | 688.465 | 2.06690 | 4.45301 | 9.59372 | .113250 |
| 8.84 | 78.1456 | 2.97321 | 9.40213 | 690.807 | 2.06768 | 4.45469 | 9.59734 | .113122 |
| 8.85 | 78.3225 | 2.97489 | 9.40744 | 693.154 | 2.06846 | 4.45637 | 9.60095 | .112994 |
| 8.86 | 78.4996 | 2.97658 | 9.41276 | 695.506 | 2.06924 | 4.45805 | 9.60457 | .112867 |
| 8.87 | 78.6769 | 2.97825 | 9.41807 | 697.864 | 2.07002 | 4.45972 | 9.60818 | .112740 |
| 8.88 | 78.8544 | 2.97993 | 9.42338 | 700.227 | 2.07080 | 4.46140 | 9.61179 | .112613 |
| 8.89 | 79.0321 | 2.98161 | 9.42868 | 702.595 | 2.07157 | 4.46307 | 9.61540 | .112486 |
| **8.90** | 79.2100 | 2.98329 | 9.43398 | 704.969 | 2.07235 | 4.46475 | 9.61900 | .112360 |
| 8.91 | 79.3881 | 2.98496 | 9.43928 | 707.348 | 2.07313 | 4.46642 | 9.62260 | .112233 |
| 8.92 | 79.5664 | 2.98664 | 9.44458 | 709.732 | 2.07390 | 4.46809 | 9.62620 | .112108 |
| 8.93 | 79.7449 | 2.98831 | 9.44987 | 712.122 | 2.07468 | 4.46976 | 9.62980 | .111982 |
| 8.94 | 79.9236 | 2.98998 | 9.45516 | 714.517 | 2.07545 | 4.47142 | 9.63339 | .111857 |
| 8.95 | 80.1025 | 2.99166 | 9.46044 | 716.917 | 2.07622 | 4.47309 | 9.63698 | .111732 |
| 8.96 | 80.2816 | 2.99333 | 9.46573 | 719.323 | 2.07700 | 4.47476 | 9.64057 | .111607 |
| 8.97 | 80.4609 | 2.99500 | 9.47101 | 721.734 | 2.07777 | 4.47642 | 9.64415 | .111483 |
| 8.98 | 80.6404 | 2.99666 | 9.47629 | 724.151 | 2.07854 | 4.47808 | 9.64774 | .111359 |
| 8.99 | 80.8201 | 2.99833 | 9.48156 | 726.573 | 2.07931 | 4.47974 | 9.65132 | .111235 |
| **9.00** | 81.0000 | 3.00000 | 9.48683 | 729.000 | 2.08008 | 4.48140 | 9.65489 | .111111 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

# Powers — Roots — Reciprocals

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **9.00** | 81.0000 | 3.00000 | 9.48683 | 729.000 | 2.08008 | 4.48140 | 9.65489 | .111111 |
| 9.01 | 81.1801 | 3.00167 | 9.49210 | 731.433 | 2.08085 | 4.48306 | 9.65847 | .110988 |
| 9.02 | 81.3604 | 3.00333 | 9.49737 | 733.871 | 2.08162 | 4.48472 | 9.66204 | .110865 |
| 9.03 | 81.5409 | 3.00500 | 9.50263 | 736.314 | 2.08239 | 4.48638 | 9.66561 | .110742 |
| 9.04 | 81.7216 | 3.00666 | 9.50789 | 738.763 | 2.08316 | 4.48803 | 9.66918 | .110619 |
| 9.05 | 81.9025 | 3.00832 | 9.51315 | 741.218 | 2.08393 | 4.48969 | 9.67274 | .110497 |
| 9.06 | 82.0836 | 3.00998 | 9.51840 | 743.677 | 2.08470 | 4.49134 | 9.67630 | .110375 |
| 9.07 | 82.2649 | 3.01164 | 9.52365 | 746.143 | 2.08546 | 4.49299 | 9.67986 | .110254 |
| 9.08 | 82.4464 | 3.01330 | 9.52890 | 748.613 | 2.08623 | 4.49464 | 9.68342 | .110132 |
| 9.09 | 82.6281 | 3.01496 | 9.53415 | 751.089 | 2.08699 | 4.49629 | 9.68697 | .110011 |
| **9.10** | 82.8100 | 3.01662 | 9.53939 | 753.571 | 2.08776 | 4.49794 | 9.69052 | .109890 |
| 9.11 | 82.9921 | 3.01828 | 9.54463 | 756.058 | 2.08852 | 4.49959 | 9.69407 | .109769 |
| 9.12 | 83.1744 | 3.01993 | 9.54987 | 758.551 | 2.08929 | 4.50123 | 9.69762 | .109649 |
| 9.13 | 83.3569 | 3.02159 | 9.55510 | 761.048 | 2.09005 | 4.50288 | 9.70116 | .109529 |
| 9.14 | 83.5396 | 3.02324 | 9.56033 | 763.552 | 2.09081 | 4.50452 | 9.70470 | .109409 |
| 9.15 | 83.7225 | 3.02490 | 9.56556 | 766.061 | 2.09158 | 4.50616 | 9.70824 | .109290 |
| 9.16 | 83.9056 | 3.02655 | 9.57079 | 768.575 | 2.09234 | 4.50781 | 9.71177 | .109170 |
| 9.17 | 84.0889 | 3.02820 | 9.57601 | 771.095 | 2.09310 | 4.50945 | 9.71531 | .109051 |
| 9.18 | 84.2724 | 3.02985 | 9.58123 | 773.621 | 2.09386 | 4.51108 | 9.71884 | .108932 |
| 9.19 | 84.4561 | 3.03150 | 9.58645 | 776.152 | 2.09462 | 4.51272 | 9.72236 | .108814 |
| **9.20** | 84.6400 | 3.03315 | 9.59166 | 778.688 | 2.09538 | 4.51436 | 9.72589 | .108696 |
| 9.21 | 84.8241 | 3.03480 | 9.59687 | 781.230 | 2.09614 | 4.51599 | 9.72941 | .108578 |
| 9.22 | 85.0084 | 3.03645 | 9.60208 | 783.777 | 2.09690 | 4.51763 | 9.73293 | .108460 |
| 9.23 | 85.1929 | 3.03809 | 9.60729 | 786.330 | 2.09765 | 4.51926 | 9.73645 | .108342 |
| 9.24 | 85.3776 | 3.03974 | 9.61249 | 788.889 | 2.09841 | 4.52089 | 9.73996 | .108225 |
| 9.25 | 85.5625 | 3.04138 | 9.61769 | 791.453 | 2.09917 | 4.52252 | 9.74348 | .108108 |
| 9.26 | 85.7476 | 3.04302 | 9.62289 | 794.023 | 2.09992 | 4.52415 | 9.74699 | .107991 |
| 9.27 | 85.9329 | 3.04467 | 9.62808 | 796.598 | 2.10068 | 4.52578 | 9.75049 | .107875 |
| 9.28 | 86.1184 | 3.04631 | 9.63328 | 799.179 | 2.10144 | 4.52740 | 9.75400 | .107759 |
| 9.29 | 86.3041 | 3.04795 | 9.63846 | 801.765 | 2.10219 | 4.52903 | 9.75750 | .107643 |
| **9.30** | 86.4900 | 3.04959 | 9.64365 | 804.357 | 2.10294 | 4.53065 | 9.76100 | .107527 |
| 9.31 | 86.6761 | 3.05123 | 9.64883 | 806.954 | 2.10370 | 4.53228 | 9.76450 | .107411 |
| 9.32 | 86.8624 | 3.05287 | 9.65401 | 809.558 | 2.10445 | 4.53390 | 9.76799 | .107296 |
| 9.33 | 87.0489 | 3.05450 | 9.65919 | 812.166 | 2.10520 | 4.53552 | 9.77148 | .107181 |
| 9.34 | 87.2356 | 3.05614 | 9.66437 | 814.781 | 2.10595 | 4.53714 | 9.77497 | .107066 |
| 9.35 | 87.4225 | 3.05778 | 9.66954 | 817.400 | 2.10671 | 4.53876 | 9.77846 | .106952 |
| 9.36 | 87.6096 | 3.05941 | 9.67471 | 820.026 | 2.10746 | 4.54038 | 9.78195 | .106838 |
| 9.37 | 87.7969 | 3.06105 | 9.67988 | 822.657 | 2.10821 | 4.54199 | 9.78543 | .106724 |
| 9.38 | 87.9844 | 3.06268 | 9.68504 | 825.294 | 2.10896 | 4.54361 | 9.78891 | .106610 |
| 9.39 | 88.1721 | 3.06431 | 9.69020 | 827.936 | 2.10971 | 4.54522 | 9.79239 | .106496 |
| **9.40** | 88.3600 | 3.06594 | 9.69536 | 830.584 | 2.11045 | 4.54684 | 9.79586 | .106383 |
| 9.41 | 88.5481 | 3.06757 | 9.70052 | 833.238 | 2.11120 | 4.54845 | 9.79933 | .106270 |
| 9.42 | 88.7364 | 3.06920 | 9.70567 | 835.897 | 2.11195 | 4.55006 | 9.80280 | .106157 |
| 9.43 | 88.9249 | 3.07083 | 9.71082 | 838.562 | 2.11270 | 4.55167 | 9.80627 | .106045 |
| 9.44 | 89.1136 | 3.07246 | 9.71597 | 841.232 | 2.11344 | 4.55328 | 9.80974 | .105932 |
| 9.45 | 89.3025 | 3.07409 | 9.72111 | 843.909 | 2.11419 | 4.55488 | 9.81320 | .105820 |
| 9.46 | 89.4916 | 3.07571 | 9.72625 | 846.591 | 2.11494 | 4.55649 | 9.81666 | .105708 |
| 9.47 | 89.6809 | 3.07734 | 9.73139 | 849.278 | 2.11568 | 4.55809 | 9.82012 | .105597 |
| 9.48 | 89.8704 | 3.07896 | 9.73653 | 851.971 | 2.11642 | 4.55970 | 9.82357 | .105485 |
| 9.49 | 90.0601 | 3.08058 | 9.74166 | 854.670 | 2.11717 | 4.56130 | 9.82703 | .105374 |
| **9.50** | 90.2500 | 3.08221 | 9.74679 | 857.375 | 2.11791 | 4.56290 | 9.83048 | .105263 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |
|---|---|---|---|---|---|---|---|---|
| **9.50** | 90.2500 | 3.08221 | 9.74679 | 857.375 | 2.11791 | 4.56290 | 9.83048 | .105263 |
| 9.51 | 90.4401 | 3.08383 | 9.75192 | 860.085 | 2.11865 | 4.56450 | 9.83392 | .105152 |
| 9.52 | 90.6304 | 3.08545 | 9.75705 | 862.801 | 2.11940 | 4.56610 | 9.83737 | .105042 |
| 9.53 | 90.8209 | 3.08707 | 9.76217 | 865.523 | 2.12014 | 4.56770 | 9.84081 | .104932 |
| 9.54 | 91.0116 | 3.08869 | 9.76729 | 868.251 | 2.12088 | 4.56930 | 9.84425 | .104822 |
| 9.55 | 91.2025 | 3.09031 | 9.77241 | 870.984 | 2.12162 | 4.57089 | 9.84769 | .104712 |
| 9.56 | 91.3936 | 3.09192 | 9.77753 | 873.723 | 2.12236 | 4.57249 | 9.85113 | .104603 |
| 9.57 | 91.5849 | 3.09354 | 9.78264 | 876.467 | 2.12310 | 4.57408 | 9.85456 | .104493 |
| 9.58 | 91.7764 | 3.09516 | 9.78775 | 879.218 | 2.12384 | 4.57567 | 9.85799 | .104384 |
| 9.59 | 91.9681 | 3.09677 | 9.79285 | 881.974 | 2.12458 | 4.57727 | 9.86142 | .104275 |
| **9.60** | 92.1600 | 3.09839 | 9.79796 | 884.736 | 2.12532 | 4.57886 | 9.86485 | .104167 |
| 9.61 | 92.3521 | 3.10000 | 9.80306 | 887.504 | 2.12605 | 4.58045 | 9.86827 | .104058 |
| 9.62 | 92.5444 | 3.10161 | 9.80816 | 890.277 | 2.12679 | 4.58204 | 9.87169 | .103950 |
| 9.63 | 92.7369 | 3.10322 | 9.81326 | 893.056 | 2.12753 | 4.58362 | 9.87511 | .103842 |
| 9.64 | 92.9296 | 3.10483 | 9.81835 | 895.841 | 2.12826 | 4.58521 | 9.87853 | .103734 |
| 9.65 | 93.1225 | 3.10644 | 9.82344 | 898.632 | 2.12900 | 4.58679 | 9.88195 | .103627 |
| 9.66 | 93.3156 | 3.10805 | 9.82853 | 901.429 | 2.12974 | 4.58838 | 9.88536 | .103520 |
| 9.67 | 93.5089 | 3.10966 | 9.83362 | 904.231 | 2.13047 | 4.58996 | 9.88877 | .103413 |
| 9.68 | 93.7024 | 3.11127 | 9.83870 | 907.039 | 2.13120 | 4.59154 | 9.89217 | .103306 |
| 9.69 | 93.8961 | 3.11288 | 9.84378 | 909.853 | 2.13194 | 4.59312 | 9.89558 | .103199 |
| **9.70** | 94.0900 | 3.11448 | 9.84886 | 912.673 | 2.13267 | 4.59470 | 9.89898 | .103093 |
| 9.71 | 94.2841 | 3.11609 | 9.85393 | 915.499 | 2.13340 | 4.59628 | 9.90238 | .102987 |
| 9.72 | 94.4784 | 3.11769 | 9.85901 | 918.330 | 2.13414 | 4.59786 | 9.90578 | .102881 |
| 9.73 | 94.6729 | 3.11929 | 9.86408 | 921.167 | 2.13487 | 4.59943 | 9.90918 | .102775 |
| 9.74 | 94.8676 | 3.12090 | 9.86914 | 924.010 | 2.13560 | 4.60101 | 9.91257 | .102669 |
| 9.75 | 95.0625 | 3.12250 | 9.87421 | 926.859 | 2.13633 | 4.60258 | 9.91596 | .102564 |
| 9.76 | 95.2576 | 3.12410 | 9.87927 | 929.714 | 2.13706 | 4.60416 | 9.91935 | .102459 |
| 9.77 | 95.4529 | 3.12570 | 9.88433 | 932.575 | 2.13779 | 4.60573 | 9 92274 | .102354 |
| 9.78 | 95.6484 | 3.12730 | 9.88939 | 935.441 | 2.13852 | 4.60730 | 9.92612 | .102249 |
| 9.79 | 95.8441 | 3.12890 | 9.89444 | 938.314 | 2.13925 | 4.60887 | 9.92950 | .102145 |
| **9.80** | 96.0400 | 3.13050 | 9.89949 | 941.192 | 2.13997 | 4.61044 | 9.93288 | .102041 |
| 9.81 | 96.2361 | 3.13209 | 9.90454 | 944.076 | 2.14070 | 4.61200 | 9.93626 | .101937 |
| 9.82 | 96.4324 | 3.13369 | 9.90959 | 946.966 | 2.14143 | 4.61357 | 9.93964 | .101833 |
| 9.83 | 96.6289 | 3.13528 | 9.91464 | 949.862 | 2.14216 | 4 61514 | 9.94301 | .101729 |
| 9.84 | 96.8256 | 3.13688 | 9.91968 | 952.764 | 2.14288 | 4.61670 | 9.94638 | .101626 |
| 9.85 | 97.0225 | 3.13847 | 9.92472 | 955.672 | 2.14361 | 4.61826 | 9.94975 | .101523 |
| 9.86 | 97.2196 | 3.14006 | 9.92975 | 958.585 | 2.14433 | 4.61983 | 9.95311 | .101420 |
| 9.87 | 97.4169 | 3.14166 | 9.93479 | 961.505 | 2.14506 | 4.62139 | 9.95648 | .101317 |
| 9.88 | 97.6144 | 3.14325 | 9.93982 | 964.430 | 2.14578 | 4.62295 | 9.95984 | .101215 |
| 9.89 | 97.8121 | 3.14484 | 9.94485 | 967.362 | 2.14651 | 4.62451 | 9.96320 | .101112 |
| **9.90** | 98.0100 | 3.14643 | 9.94987 | 970.299 | 2.14723 | 4.62607 | 9.96655 | .101010 |
| 9.91 | 98.2081 | 3.14802 | 9.95490 | 973.242 | 2.14795 | 4.62762 | 9.96991 | .100908 |
| 9.92 | 98.4064 | 3.14960 | 9.95992 | 976.191 | 2.14867 | 4.62918 | 9.97326 | .100806 |
| 9.93 | 98.6049 | 3.15119 | 9.96494 | 979.147 | 2.14940 | 4.63073 | 9.97661 | .100705 |
| 9.94 | 98.8036 | 3.15278 | 9.96995 | 982.108 | 2.15012 | 4.63229 | 9.97996 | .100604 |
| 9.95 | 99.0025 | 3.15436 | 9.97497 | 985.075 | 2.15084 | 4.63384 | 9.98331 | .100503 |
| 9.96 | 99.2016 | 3.15595 | 9.97998 | 988.048 | 2.15156 | 4.63539 | 9.98665 | .100402 |
| 9.97 | 99.4009 | 3.15753 | 9.98499 | 991.027 | 2.15228 | 4.63694 | 9.98999 | .100301 |
| 9.98 | 99.6004 | 3.15911 | 9.98999 | 994.012 | 2.15300 | 4.63849 | 9.99333 | .100200 |
| 9.99 | 99.8001 | 3.16070 | 9.99500 | 997.003 | 2.15372 | 4.64004 | 9.99667 | .100100 |
| **10.00** | 100.000 | 3.16228 | 10.0000 | 1000.00 | 2.15443 | 4.64159 | 10.0000 | .100000 |
| $n$ | $n^2$ | $\sqrt{n}$ | $\sqrt{10n}$ | $n^3$ | $\sqrt[3]{n}$ | $\sqrt[3]{10n}$ | $\sqrt[3]{100n}$ | $1/n$ |

| N | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 2 3 | 4 5 6 | 7 8 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **10** | 0000 | 0043 | 0086 | 0128 | 0170 | 0212 | 0253 | 0294 | 0334 | 0374 | 4 8 12 | 17 21 25 | 29 33 37 |
| 11 | 0414 | 0453 | 0492 | 0531 | 0569 | 0607 | 0645 | 0682 | 0719 | 0755 | 4 8 11 | 15 19 23 | 26 30 34 |
| 12 | 0792 | 0828 | 0864 | 0899 | 0934 | 0969 | 1004 | 1038 | 1072 | 1106 | 3 7 10 | 14 17 21 | 24 28 31 |
| 13 | 1139 | 1173 | 1206 | 1239 | 1271 | 1303 | 1335 | 1367 | 1399 | 1430 | 3 6 10 | 13 16 19 | 23 26 29 |
| 14 | 1461 | 1492 | 1523 | 1553 | 1584 | 1614 | 1644 | 1673 | 1703 | 1732 | 3 6 9 | 12 15 18 | 21 24 27 |
| **15** | 1761 | 1790 | 1818 | 1847 | 1875 | 1903 | 1931 | 1959 | 1987 | 2014 | 3 6 8 | 11 14 17 | 20 22 25 |
| 16 | 2041 | 2068 | 2095 | 2122 | 2148 | 2175 | 2201 | 2227 | 2253 | 2279 | 3 5 8 | 11 13 16 | 18 21 24 |
| 17 | 2304 | 2330 | 2355 | 2380 | 2405 | 2430 | 2455 | 2480 | 2504 | 2529 | 2 5 7 | 10 12 15 | 17 20 22 |
| 18 | 2553 | 2577 | 2601 | 2625 | 2648 | 2672 | 2695 | 2718 | 2742 | 2765 | 2 5 7 | 9 12 14 | 16 19 21 |
| 19 | 2788 | 2810 | 2833 | 2856 | 2878 | 2900 | 2923 | 2945 | 2967 | 2989 | 2 4 7 | 9 11 13 | 16 18 20 |
| **20** | 3010 | 3032 | 3054 | 3075 | 3096 | 3118 | 3139 | 3160 | 3181 | 3201 | 2 4 6 | 8 11 13 | 15 17 19 |
| 21 | 3222 | 3243 | 3263 | 3284 | 3304 | 3324 | 3345 | 3365 | 3385 | 3404 | 2 4 6 | 8 10 12 | 14 16 18 |
| 22 | 3424 | 3444 | 3464 | 3483 | 3502 | 3522 | 3541 | 3560 | 3579 | 3598 | 2 4 6 | 8 10 12 | 14 16 17 |
| 23 | 3617 | 3636 | 3655 | 3674 | 3692 | 3711 | 3729 | 3747 | 3766 | 3784 | 2 4 6 | 7 9 11 | 13 15 17 |
| 24 | 3802 | 3820 | 3838 | 3856 | 3874 | 3892 | 3909 | 3927 | 3945 | 3962 | 2 4 5 | 7 9 11 | 12 14 16 |
| **25** | *3979* | *3997* | *4014* | *4031* | *4048* | *4065* | *4082* | *4099* | *4116* | *4133* | *2 4 5* | *7 9 10* | *12 14 16* |
| 26 | 4150 | 4166 | 4183 | 4200 | 4216 | 4232 | 4249 | 4265 | 4281 | 4298 | 2 3 5 | 7 8 10 | 11 13 15 |
| 27 | 4314 | 4330 | 4346 | 4362 | 4378 | 4393 | 4409 | 4425 | 4440 | 4456 | 2 3 5 | 6 8 9 | 11 12 14 |
| 28 | 4472 | 4487 | 4502 | 4518 | 4533 | 4548 | 4564 | 4579 | 4594 | 4609 | 2 3 5 | 6 8 9 | 11 12 14 |
| 29 | 4624 | 4639 | 4654 | 4669 | 4683 | 4698 | 4713 | 4728 | 4742 | 4757 | 1 3 4 | 6 7 9 | 10 12 13 |
| **30** | 4771 | 4786 | 4800 | 4814 | 4829 | 4843 | 4857 | 4871 | 4886 | 4900 | 1 3 4 | 6 7 9 | 10 11 13 |
| 31 | 4914 | 4928 | 4942 | 4955 | 4969 | 4983 | 4997 | 5011 | 5024 | 5038 | 1 3 4 | 5 7 8 | 10 11 12 |
| 32 | 5051 | 5065 | 5079 | 5092 | 5105 | 5119 | 5132 | 5145 | 5159 | 5172 | 1 3 4 | 5 7 8 | 9 11 12 |
| 33 | *5185* | *5198* | *5211* | *5224* | *5237* | *5250* | *5263* | *5276* | *5289* | *5302* | *1 3 4* | *5 7 8* | *9 11 12* |
| 34 | 5315 | 5328 | 5340 | 5353 | 5366 | 5378 | 5391 | 5403 | 5416 | 5428 | 1 2 4 | 5 6 8 | 9 10 11 |
| **35** | 5441 | 5453 | 5465 | 5478 | 5490 | 5502 | 5514 | 5527 | 5539 | 5551 | 1 2 4 | 5 6 7 | 9 10 11 |
| 36 | 5563 | 5575 | 5587 | 5599 | 5611 | 5623 | 5635 | 5647 | 5658 | 5670 | 1 2 4 | 5 6 7 | 8 10 11 |
| 37 | 5682 | 5694 | 5705 | 5717 | 5729 | 5740 | 5752 | 5763 | 5775 | 5786 | 1 2 4 | 5 6 7 | 8 9 11 |
| 38 | 5798 | 5809 | 5821 | 5832 | 5843 | 5855 | 5866 | 5877 | 5888 | 5899 | 1 2 3 | 5 6 7 | 8 9 10 |
| 39 | 5911 | 5922 | 5933 | 5944 | 5955 | 5966 | 5977 | 5988 | 5999 | 6010 | 1 2 3 | 4 5 7 | 8 9 10 |
| **40** | 6021 | 6031 | 6042 | 6053 | 6064 | 6075 | 6085 | 6096 | 6107 | 6117 | 1 2 3 | 4 5 6 | 8 9 10 |
| 41 | 6128 | 6138 | 6149 | 6160 | 6170 | 6180 | 6191 | 6201 | 6212 | 6222 | 1 2 3 | 4 5 6 | 7 8 9 |
| 42 | 6232 | 6243 | 6253 | 6263 | 6274 | 6284 | 6294 | 6304 | 6314 | 6325 | 1 2 3 | 4 5 6 | 7 8 9 |
| 43 | 6335 | 6345 | 6355 | 6365 | 6375 | 6385 | 6395 | 6405 | 6415 | 6425 | 1 2 3 | 4 5 6 | 7 8 9 |
| 44 | 6435 | 6444 | 6454 | 6464 | 6474 | 6484 | 6493 | 6503 | 6513 | 6522 | 1 2 3 | 4 5 6 | 7 8 9 |
| **45** | 6532 | 6542 | 6551 | 6561 | 6571 | 6580 | 6590 | 6599 | 6609 | 6618 | 1 2 3 | 4 5 6 | 7 8 9 |
| 46 | 6628 | 6637 | 6646 | 6656 | 6665 | 6675 | 6684 | 6693 | 6702 | 6712 | 1 2 3 | 4 5 6 | 7 7 8 |
| 47 | 6721 | 6730 | 6739 | 6749 | 6758 | 6767 | 6776 | 6785 | 6794 | 6803 | 1 2 3 | 4 5 6 | 7 7 8 |
| 48 | 6812 | 6821 | 6830 | 6839 | 6848 | 6857 | 6866 | 6875 | 6884 | 6893 | 1 2 3 | 4 5 6 | 7 7 8 |
| 49 | *6902* | *6911* | *6920* | *6928* | *6937* | *6946* | *6955* | *6964* | *6972* | *6981* | 1 2 3 | 4 4 5 | 6 7 8 |
| **50** | 6990 | 6998 | 7007 | 7016 | 7024 | 7033 | 7042 | 7050 | 7059 | 7067 | 1 2 3 | 3 4 5 | 6 7 8 |
| 51 | 7076 | 7084 | 7093 | 7101 | 7110 | 7118 | 7126 | 7135 | 7143 | 7152 | 1 2 3 | 3 4 5 | 6 7 8 |
| 52 | 7160 | 7168 | 7177 | 7185 | 7193 | 7202 | 7210 | 7218 | 7226 | 7235 | 1 2 3 | 3 4 5 | 6 7 7 |
| 53 | 7243 | 7251 | 7259 | 7267 | 7275 | 7284 | 7292 | 7300 | 7308 | 7316 | 1 2 2 | 3 4 5 | 6 6 7 |
| 54 | 7324 | 7332 | 7340 | 7348 | 7356 | 7364 | 7372 | 7380 | 7388 | 7396 | 1 2 2 | 3 4 5 | 6 6 7 |
| **N** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | 1 2 2 | 4 5 6 | 7 8 9 |

**The proportional parts are stated in full for every tenth at the right-hand side. The logarithm of any number of four significant figures can be read directly by add-**

| N | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **55** | 7404 | 7412 | 7419 | 7427 | 7435 | 7443 | 7451 | 7459 | 7466 | 7474 | 1 | 2 | 2 | 3 | 4 | 5 | 5 | 6 | 7 |
| 56 | 7482 | 7490 | 7497 | 7505 | 7513 | 7520 | 7528 | 7536 | 7543 | 7551 | 1 | 2 | 2 | 3 | 4 | 5 | 5 | 6 | 7 |
| 57 | 7559 | 7566 | 7574 | 7582 | 7589 | 7597 | 7604 | 7612 | 7619 | 7627 | 1 | 1 | 2 | 3 | 4 | 5 | 5 | 6 | 7 |
| 58 | 7634 | 7642 | 7649 | 7657 | 7664 | 7672 | 7679 | 7686 | 7694 | 7701 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 7 |
| 59 | 7709 | 7716 | 7723 | 7731 | 7738 | 7745 | 7752 | 7760 | 7767 | 7774 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 7 |
| **60** | 7782 | 7789 | 7796 | 7803 | 7810 | 7818 | 7825 | 7832 | 7839 | 7846 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 6 |
| 61 | 7853 | 7860 | 7868 | 7875 | 7882 | 7889 | 7896 | 7903 | 7910 | 7917 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 6 | 6 |
| 62 | 7924 | 7931 | 7938 | 7945 | 7952 | 7959 | 7966 | 7973 | 7980 | 7987 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 5 | 6 |
| 63 | 7993 | 8000 | 8007 | 8014 | 8021 | 8028 | 8035 | 8041 | 8048 | 8055 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 5 | 6 |
| 64 | 8062 | 8069 | 8075 | 8082 | 8089 | 8096 | 8102 | 8109 | 8116 | 8122 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 5 | 6 |
| **65** | 8129 | 8136 | 8142 | 8149 | 8156 | 8162 | 8169 | 8176 | 8182 | 8189 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 5 | 6 |
| 66 | 8195 | 8202 | 8209 | 8215 | 8222 | 8228 | 8235 | 8241 | 8248 | 8254 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 5 | 6 |
| 67 | 8261 | 8267 | 8274 | 8280 | 8287 | 8293 | 8299 | 8306 | 8312 | 8319 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 5 | 6 |
| 68 | 8325 | 8331 | 8338 | 8344 | 8351 | 8357 | 8363 | 8370 | 8376 | 8382 | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 6 |
| 69 | 8388 | 8395 | 8401 | 8407 | 8414 | 8420 | 8426 | 8432 | 8439 | 8445 | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 6 |
| **70** | 8451 | 8457 | 8463 | 8470 | 8476 | 8482 | 8488 | 8494 | 8500 | 8506 | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 6 |
| 71 | 8513 | 8519 | 8525 | 8531 | 8537 | 8543 | 8549 | 8555 | 8561 | 8567 | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 6 |
| 72 | 8573 | 8579 | 8585 | 8591 | 8597 | 8603 | 8609 | 8615 | 8621 | 8627 | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 6 |
| 73 | 8633 | 8639 | 8645 | 8651 | 8657 | 8663 | 8669 | 8675 | 8681 | 8686 | 1 | 1 | 2 | 2 | 3 | 4 | 4 | 5 | 5 |
| 74 | 8692 | 8698 | 8704 | 8710 | 8716 | 8722 | 8727 | 8733 | 8739 | 8745 | 1 | 1 | 2 | 2 | 3 | 4 | 4 | 5 | 5 |
| **75** | 8751 | 8756 | 8762 | 8768 | 8774 | 8779 | 8785 | 8791 | 8797 | 8802 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 5 | 5 |
| 76 | 8808 | 8814 | 8820 | 8825 | 8831 | 8837 | 8842 | 8848 | 8854 | 8859 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 77 | 8865 | 8871 | 8876 | 8882 | 8887 | 8893 | 8899 | 8904 | 8910 | 8915 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 78 | 8921 | 8927 | 8932 | 8938 | 8943 | 8949 | 8954 | 8960 | 8965 | 8971 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 79 | 8976 | 8982 | 8987 | 8993 | 8998 | 9004 | 9009 | 9015 | 9020 | 9025 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| **80** | 9031 | 9036 | 9042 | 9047 | 9053 | 9058 | 9063 | 9069 | 9074 | 9079 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 81 | 9085 | 9090 | 9096 | 9101 | 9106 | 9112 | 9117 | 9122 | 9128 | 9133 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 82 | 9138 | 9143 | 9149 | 9154 | 9159 | 9165 | 9170 | 9175 | 9180 | 9186 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 83 | 9191 | 9196 | 9201 | 9206 | 9212 | 9217 | 9222 | 9227 | 9232 | 9238 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 84 | 9243 | 9248 | 9253 | 9258 | 9263 | 9269 | 9274 | 9279 | 9284 | 9289 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| **85** | 9294 | 9299 | 9304 | 9309 | 9315 | 9320 | 9325 | 9330 | 9335 | 9340 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 86 | 9345 | 9350 | 9355 | 9360 | 9365 | 9370 | 9375 | 9380 | 9385 | 9390 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 87 | 9395 | 9400 | 9405 | 9410 | 9415 | 9420 | 9425 | 9430 | 9435 | 9440 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| 88 | 9445 | 9450 | 9455 | 9460 | 9465 | 9469 | 9474 | 9479 | 9484 | 9489 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 89 | 9494 | 9499 | 9504 | 9509 | 9513 | 9518 | 9523 | 9528 | 9533 | 9538 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| **90** | 9542 | 9547 | 9552 | 9557 | 9562 | 9566 | 9571 | 9576 | 9581 | 9586 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 91 | 9590 | 9595 | 9600 | 9605 | 9609 | 9614 | 9619 | 9624 | 9628 | 9633 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 92 | 9638 | 9643 | 9647 | 9652 | 9657 | 9661 | 9666 | 9671 | 9675 | 9680 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 93 | 9685 | 9689 | 9694 | 9699 | 9703 | 9708 | 9713 | 9717 | 9722 | 9727 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 94 | 9731 | 9736 | 9741 | 9745 | 9750 | 9754 | 9759 | 9763 | 9768 | 9773 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| **95** | 9777 | 9782 | 9786 | 9791 | 9795 | 9800 | 9805 | 9809 | 9814 | 9818 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 96 | 9823 | 9827 | 9832 | 9836 | 9841 | 9845 | 9850 | 9854 | 9859 | 9863 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 97 | 9868 | 9872 | 9877 | 9881 | 9886 | 9890 | 9894 | 9899 | 9903 | 9908 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 98 | 9912 | 9917 | 9921 | 9926 | 9930 | 9934 | 9939 | 9943 | 9948 | 9952 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 4 |
| 99 | 9956 | 9961 | 9965 | 9969 | 9974 | 9978 | 9983 | 9987 | 9991 | 9996 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 4 |
| N | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

ing the proportional part *corresponding to* the *fourth figure to* the tabular number corresponding to the first three figures. There may be an error of 1 in the last place.

# SUBJECT INDEX

(The citations are to pages and include only the subject matter. A *Personal Index* is appended separately.)


Abscissa: continuous frequency series and units on the, 227; definition of, 219; equal distances on, represent equal facts, 222; meaning of discrete units on the, 222-223; reading of units on the, with cumulations on a "less than" basis, 237, on a "more than" basis, 237; relation of the, to ordinate scale in plotting time series, 244-245; units on the, points rather than spaces, 223-225; units on the, in time series, 243-244; zero position on the, 221.

Accounting: contrasted with Statistics, 1-2.

Accuracy: comparability of data and, 42-44; error and, 40; estimates and, 41; secondary data and, 38-42.

Accuracy of determination, 41.

Accuracy with which facts are determined, 39-40.

Accuracy with which facts are reported, 39.

Annalist (*See* Index numbers, barometric and forecasting).

*Annalist's Index Number* (*See* Index numbers, wholesale prices.)

Arithmetic mean: algebraic sum of deviations from the, 271, 337; as the "true" value, 266; compensatory effects of weights in the computation of the, 268-269, 271; definition of the, 263; distribution of errors from the, 266; effect on the, of an arbitrary selection of weights, 269-270; explanation of the, 264-267; "fictitious" character of an, 265; how computed, 267-278; illustration of the, as the center of gravity, 268-269; properties of the, 310-311; "real" character of the, 265; "short-cut" method of calculating the, 271-278, from an assumed average, 271-275, from an assumed average by "steps," 276-278; summary of properties of the, 282; use of the, as substitute for detail, 265-266, in summarizing link-relatives, 451, in measurements of the physical sciences, 266, as a type in measuring dispersion, 337-338. (*See* also Averages.)

Arithmetical triangle: binomial expansion and the, 364-365; detail of the, to eleventh line, 365.

Assembling data of record, 55.

Asymmetry (*See* Skewness)..

Average: an, as a statistical constant, 262, a substitute for variables, 262; conditions upon which the use of an, depends, 263; the nature of an, 320, 322; as a type in measuring dispersion, 337; type to use, in averaging deviations, 338-339.

Average deviation: calculation of the, in grouped series from an assumed average, 346, from an assumed average by the "step" method, 346-348, from the true average, 345; calculation of the, in historical series from an assumed average, 341-342, from a true average, 339-341; calculation of the, in ungrouped frequency series from the true average, 344; definition of the, 337; nature of the, 337; relation of the, to the standard deviation, 351, 355. (*See* Dispersion.)

Average of relatives (*See* Index numbers, general).

Averages, simple: as derivative expressions, 321; as types, 261-323; cautions necessary in using, 319-320; common, defined, 263-264; differences between simple and weighted, 279-281; do's and don'ts in the use of, 278-281; "first" order of, 264-324; loose use of, 262-263; properties of, 320; purpose of "second order" of, 324; relation of, to purpose and use, 321; results of averaging, 278-279; the, to use in typical cases, 311-315; use of, in measuring normal seasonal variation, 449; zero cases included in calculating, 281. (*See* Arithmetic mean, Geometric mean, Median, Mode.)

Averages, weighted: simple averages contrasted with, 279-281. (*See* Index numbers, general.)

Averaging: effect of, non-homogeneous data, 315-318; meaning of the process of, in continuous series, 321, in discrete series, 321.

Axes, co-ordinate: relations between the, in graphic presentation of frequency series, 221.

Axis, abscissa (*See* Abscissa).

Axis, ordinate (*See* Ordinate).

Babson Statistical Service (*See* Index numbers, barometric and forecasting).

Base shifting: averages of relatives and, 500-502; geometric mean and, in index numbers, 501-502; medians of relatives and, in index numbers, 497-498, 501; methods of, with arithmetic means of relatives, 500-501; ratios of averages and, in index numbers, 503; "short - cut" method of, with arithmetic means of relatives, 500-501; weighted aggregate of actual prices and, 504-505. (*See* Index numbers, general.)

Bias: Fisher's index number formula to remove, 509-511; secondary data and, 32; types of, 32.

Binomial expansion: arithmetical triangle and the, 364-365; tossing of coins and the, 361-365.

Bradstreet's index number (*See* Index numbers, wholesale prices).

Brookmire (*See* Index numbers, forecast of stock prices).

Bureau of Labor Statistics (*See* Index numbers, general; Index numbers, retail; Index numbers, wholesale).

Business cycle: characteristics of the, 440; use of statistics in study of the, 18. (*See* Index numbers, commodity price, of business cycles.)

Captions: contents of, in tables, 128-129.

Cartograms: use of, for illustrating statistical series, 201-212. (*See* Diagrams; Maps, statistical.)

Causation: contrast of, with correlation, 394-398.

Cause-and-effect: meaning of, 395-398.

Chain-relatives: method of calculating, 451-452. (*See* Index numbers, general.)

Chance: illustrations of the operation of, 361-364.

Circles: uses of, as diagram, 182-185, 193-195. (*See* Diagrams, Diagrammatic Presentation.)

Classification: common characteristics of data and, 127; co-ordinate and subordinate characteristics and, 128; creative nature of, 128; methods of, illustrated, 127-128; nature of, 126-128; order of processes in, 127-128; process of, for experimental purposes, 130, for fixed tabulation forms, 129-130, illustrated, 130; relation of, to tabulation, 129; routine nature of, 128; tabulation and diagrammatic presentation contrasted with, 171-173. (*See* Tabulation.)

Coefficient of correlation: formula of the, in adjusted time series,

461; lagging adjusted time series and the, 461-463; meaning of the probable error of the, 429; probable error of the, in time series, 464-465; test of "significance" of the, 429; use of the, in adjusted time series, 457-458. (*See* Correlation.)
Coefficient of dispersion (*See* Dispersion).
Coefficient of skewness (*See* Skewness).
Coefficients: corrected, 83; crude, 83; list of corrected, 89; list of crude, 89; illustrations of, relating to *condition*, 82-84, *to* space, 81, to time, 80-81; regression, meaning of, 426-428; rules for forming, 37; tests of good, 83; units as, defined, 36-37. (*See* Lines of "best fit," Regression lines.)
Collecting data: counting as a method of, 57-59; purpose and plan in, 54; standards for, 22; of record, 54-57.
Collecting primary data: conditions preliminary to, 46-53; general aspects of, 47-71; *methods* of, 54-62; use of form letters in, 65-66; use of interviews in, 65.
Collection process: descriptively considered, 54-62; functionally considered, 61-65.
Comparability of data: and accuracy, 42-44.
Comparison: contrast of, with causation and correlation, 394-398.
Competition: use of statistics in study of, 19.
Component parts: *use of circles to* show, 182-185, 193-195; use of one-dimensional diagrams to show, 179-181, 192; use of two-dimensional diagrams to show, 181-182, 197-198. (*See* Diagrams.)
Composite unit: definition and illustrations of, 36, 78. (*See* Units.)
Concurrent deviation: formula for the, method of measuring correlation, 432; measurement of correlation by the, method of, 430-432. (*See* Correlation.)
*Condition series* (*See Frequency* series, Series).
Consistency of data, 51.
Consumption, use of statistics in the study of, 19.
Continuous series: arbitrary nature of grouping in, 216; characteristics of, 162, 166-167; continuous line necessary to show, 216; contrast of, with discrete series, 162, 166; diagrams unsuited to illustrate, 215; distribution of items in, 166-167; *graphic presentation of cumulated*, 237-242; grouping of data in, 231-232; illustrations of, 214-215; interpolation for the median in, 286-289; location of the mode in, 301-304; process of averaging, meaning of, 321; rules for grouping items in, 167; smoothed lines connecting successive ordinates and, 227-229; treatment of lines connecting successive ordinates in cumulated, 239-242; widening the groups *in*, 166-167. (*See* Frequency series, Series.)
Continuous space series: diagrams unsuited in showing, 218. (*See* Series.)
Continuous time series: diagrams unsuited in showing, 216-218; problems in graphic presentation of, 243-259. (*See* Series.)
Co-ordinate axes: relations between, in graphic presentation of frequency series, 221.
Co-ordinates: system of rectangular, *illustrated*, 219. (*See* Graphic presentation.)
Correlation: assumptions underlying the Pearsonian coefficient of, 406-410; calculation of the Pearsonian coefficient of, in grouped series from an assumed mean, illustrated, 421-424, from the true mean, illustrated, 417-421; calculation of the Pearsonian coefficient of, in ungrouped series, illustrated, 413-417; cause-and-effect relations and, 395-399;

conditions indicating negative, 411, *positive*, 410-411, *zero*, 411; contrast of, with causation, 394-398; definition and explanation of, 398-400; formula of the Pearsonian coefficient of, 410-413, in adjusted time series, 461; illustrations of, by throws of dice, 400-405; meaning of, 398-405; measurement of, 406-433, by concurrent deviation method, 430-433; nature of distributions in double frequency tables indicating, 412-413; summary of discussion of, 434-436; the "sum-product" measure of, 406-430; theory and measurement of, 393-437; time series and, 457-464.
Correlation coefficient: lagging adjusted time series and the, 461-463; probable error of the, 429, in time series, 464-465; test of "significance" of the, 429; use of the, in unadjusted time series, 457-458. (*See* Correlation, Time series.)
Correlation table: illustrations of a, 418-419, 422-423, 431. (*See* Correlation.)
Cost units, 79-80.
Costs: use of statistics in the study of, 19.
Counting: as a method of collecting data, 57-59.
Cumulation: "less than" basis of, described, 233-234, illustrated, 235; lines connecting successive ordinates in discrete frequency series and, 236; "more than" basis of, described, 234, illustrated, 236; process of, 233; scale reading and "less than" basis of, 237, and "more than" basis of, 237; time series and, 258-259. (*See* Diagrammatic presentation, Diagrams, Graphic presentation.)
Curve smoothing: method of, in continuous frequency series, 229-232; purpose of, 225-227; use of, in discrete series, 225. (*See* Diagrammatic presentation, Graphic presentation.)
Cycle per cents: meaning of, 455. (*See* Time series.)
Cyclical fluctuations: explanation of, in time series, 440; method of reducing, to a common denominator, 455-457; nature of, 455-457. (*See* Time series.)

Data: availability of, 50; bias of secondary, 32; characteristics of, 125-126, expressed in a series, 126; characteristics of secondary, 45-46; collecting, of record, 54-57; collecting primary, 47-71; collection of, descriptively considered, 54-61, functionally considered, 61-65; conditions preliminary to the collection of primary, 46-53; consistency of, 51; counting as a method of collecting, 57-59; cumulative characteristics of, 126; definition of primary, 25; effect of averaging non-homogeneous, 315-318; estimates as statistical, 60; form of non-tabulated, 129; homogeneity of secondary, 42-44; methods of collecting, 54-62; nature of secondary, 31-35; need for current, 51-52; purpose of issue of secondary, 31; purpose and plan in collection of, 54; relations of the characteristics of a body of, 125; restrictions upon the use of, 52; sanction for collecting primary, 52-53; secondary, as germane to a problem, 44-46; secondary, as samples, 32; source of, and the collection process, 61-64; sources of secondary, 30-31; tests to be applied to secondary, 30-46; type of consumers of secondary, 31; types of problems requiring primary, 49-50; ways of collecting primary, 65-68. (*See* Primary data, Secondary data.)
Deciles: formulæ for locating, 331; graphic measure of dispersion based on, 333, 334-335; measure of dispersion based on, 332. (*See* Average deviation, Dispersion.)
Demand: use of statistics in study of, 19.
Deviation, average (*See* Average deviation, Dispersion).

Deviation, quartile (*See* Quartile deviation, Dispersion).
Deviation, standard (*See* Standard deviation, Dispersion).
Deviations: algebraic sum zero, taken from arithmetic mean, 337; a minimum when taken from the median, 337-338; sum of the squares, a minimum when taken from the arithmetic mean, 350.
Deviations, concurrent: measurement of correlation by the method of, 430-432. (*See* Correlation.)
Diagrammatic presentation: contrast of, with classification and tabulation, 171-173, with graphic presentation, 218-221; general aspects of, 171-213; psychology of the use of, 173-175. (*See* Diagrams, Graphic presentation.)
Diagrams: forms of, for illustrating frequency or magnitude alone, 175-200; good and bad features of alternative forms of, 175-186; rules to be observed in the use of statistical 198-200, 212-213; types of, in current use, 186-198; unsuited to illustrate continuous frequency series, 216, time series, 217-218; uses of, for illustrating frequency or magnitude in relation to spacial distribution, 201-212; uses of circles to show component parts, 182-185, 193-195; uses of one-dimensional types of, 176-177, to show component parts, 179-181, 192; uses of three-dimensional types of, 178, 186, 190; uses of two-dimensional types of, 177-178, to show component parts, 181-182, 197, 198. (*See* Diagrammatic presentation.)
Difference charts: scale adjustments in the use of, 246-248.
Difference scale: zero on a, 244.
Discrete frequency series: "graphic" presentation of cumulated, 233-237; lines connecting successive ordinates in cumulated, 236. (*See* Diagrammatic presentation, Diagrams, Tabulation.)
Discrete series: contrast of, with continuous series, 162, 166; definition of, 162; distribution of items in, 163-164; illustrations of 214-215; interpolation for the median in a, 286-289; lines connecting successive ordinates and, 223-227; location of the mode in a, 301; meaning of units on the abscissa and, 222-223; nature of measurements in, 222; plotting simple distributions describing, 222-227; process of averaging, meaning of, 321; smoothed lines connecting successive ordinates and, 225-227; widening the groups in, 163-164. (*See* Diagrammatic presentation, Diagrams, Series.)
Dispersion: assumptions regarding, based on averaging differences from a type, 336-337; coefficient of, based on the average deviation, 348-349, decile range, 329, quartiles, 357, the range, 328-329, standard deviation, 355; general meaning of, 325-326, graphic method of measuring, based on deciles, 333, 334-335; general aspects of, 324-325; illustrations of price, 490, 492, 494; measures of, 326-358; measure of, based on the average deviation, 337-349, averaging differences from a type and, 336-358, deciles, 332, the decile range, 329, quartiles, 356-358, the range of parts of series, 329-331, standard deviation, 349-358, successive deciles, 329-330; measure of, cumulative-or-moving-range, in frequency series, 328, in historical series, 327-328; measure of, the method of limits, 326-336, the range, 326-331, the range in historical series, 326-327; precise meaning of, 325-326; price, and index number making, 489-496; skewness contrasted with, 377-378; summary of measures of, 377; summary of measures and coefficients of, 358.
Distribution: form of J-shaped, 381; form of U-shaped, 380. (*See*

Continuous series, Discrete series, Frequency series, Normal law of error curve, Skewness.)
Dun's index number (*See* Index numbers, wholesale prices).

Earnings: definition of, 97.
Editing of questionnaires, 68-70. (*See* Schedules.)
Employment: use of statistics in a study of, 21.
Error: normal law of, described, 230; theory of, and number of combinations, 366-367. (*See* Chance, Continuous series, Discrete series, Frequency series, Normal law of error curve.)
Estimates: accuracy of, 41; allied material as a basis for, 59; collecting data based on, 55; direct material as a basis for, 59; statistical methods and, 60.

Fisher's index number (*See* Index numbers, wholesale prices).
Form letters: as a method of collecting primary data, 65-66.
Frequency curve: illustration of a smoothed, 231. (*See* Frequency series, Normal law of error curve, Series.)
Frequency series: calculating average deviation in, 344-348; definition of, 300; graphic presentation of, 221-242, of continuous, cumulated, 237-242; lines connecting successive ordinates in cumulated continuous, 239-242, in cumulated discrete, 236; methods of summarizing, 386-391; nature of, 158-169. (*See* Graphic presentation, Tabulation, Series.)
Frequency table: illustration of a "four-part" double, 431; treatment of groups in, 167-169. (*See* Correlation, Tabulation.)

Geometric mean: definition of, 264, 307-308; method of calculating the, 308-309; properties of the, 310-311. (*See* Averages; Index numbers, general.)
Geometric mean of relatives (*See* Index numbers, general; Index numbers, production).
Graphic presentation: abscissa units and, 222; association between variables shown by, 433-435; continuous frequency series cumulated and, 237-242; contrast of, with diagrammatic presentation, 218-221; cumulated frequency series and, 232-242; discrete frequency series cumulated and, 233-237; frequency distributions describing continuous series and, 227-232; frequency series and, 221-242; time series and, 242-259; necessity of, in illustrating continuous series, 220; ordinate scale units and, 221-222; simple frequency series and, 221-232. (*See* Diagrammatic presentation, Diagrams.)
Groups: methods of writing, 168; statistical, and discrete series, 224. (*See* Tabulation.)

Harvard Committee on Economic Research (*See* Index numbers, general business conditions).
Harvard's commodity price index of business cycles (*See* Index numbers, commodity price, of business cycles).
Historical series (*See* Time series).
Hollerith tabulation card: description of, 153-154; illustration of, 154. (*See* Tabulation.)
Homogeneity of secondary data, 42-44. (*See* Secondary data.)

Income: use of statistics in a study of, 19.
Index numbers, applicants for work, 545.
Index numbers, bank clearings, 544.
Index numbers, barometric and forecasting: Annalist's, 543; Babson's 543; Standard Statistics Corporation's, 543.
Index numbers, bond yields, 545.
Index numbers, commodity price, of business cycles, Harvard's, 529-530.
Index numbers, department store sales, 545; stocks, 545.

Index numbers, earnings and wage-rates, 545-546.
Index numbers, employment and the business cycle, 544.
Index numbers, employment in manufacturing, 544.
Index numbers, factory employment in Illinois, 545.
Index numbers, fluctuations in employment, 544.
Index numbers, forecast of stock prices, Brookmire's, 541, 543.
Index numbers, general: aggregate of actual prices, method of calculating, 477-479; aggregate of actual prices, weighted, and base shifting, 504-505, and their merits, 504; American price, explanation of, 515-530; arithmetic mean of relatives, and base shifting in, 500-501, fixed base, and, 497; attributes of price, 482-483; average of relatives method of calculating, 470-476, and base shifting, 500-502, and price dispersion with chain base, 495-496, and price dispersion with fixed base, 495-496; average of relatives, simple, method of calculating, 470-473, chain base, 472-473, fixed base, 470-472; average of relatives, weighted, method of calculating, 473-476; average of relatives with fixed versus shifting base, 496-499; bias in, and formula for constructing, 509-511; chain-relatives, and, 499-500; conclusions respecting making and using, 513; data for calculating wholesale price, 470; definition of, 469; difficulty of securing prices for, 487-488; dispersion of price fluctuations and, 489-496; dispersion of prices and the making of, 495-496; dispersion of prices, illustrated, and, 490, 492, 494, on a near base and, 490-495, on a remote base and, 490-493; effect of animal products, in price, 485, commodities included, in price, 484-486, consumers' goods, in price, 484, farm crops, in price, 485, manufactured goods, in price, 485, mineral products, in price, 485, producers' goods, in price, 484, raw products, in price, 484-485; "factor reversal" test, 511; Fisher's "Ideal" formula for, 509; general use of, 468; geometric mean of relatives and base shifting in, 501-502, fixed base, and, 498-499, method of calculating, fixed base, 472; medians of relatives, and base shifting in, 497-498, 501, fixed base, and, 497-498, method of calculating, fixed base, 471-472, weighted, method of calculating, 475-476; meaning of "price" in calculating, 486-487; methods of constructing, 496-505; number of commodities and price, 489; price, nature of, and, 469; principles of making, 481-496, and using, 468-514; processes in making, enumerated, 483-484; purpose of price, 482; quantities as weights in, 477-479; ratios of averages and base shifting in, 503; ratios of averages method of calculating, 476-477, 502; ratios of weighted aggregates method of calculating, 477-479; selection of prices and use of, 481; suggestions to users of, 511-513; summary of results of calculating, by different methods, 479-480; "time reversal" test in, 510; use of, general, 481, of price, 480-481; values as weights in, 473-474, 478-479; weighted versus unweighted series of, 505-507; weighting and Fisher's "ideal" formula in, 509-511; weighting, general aspects of, 505-511, meaning and methods of, 505-511; weighted and unweighted, similarity of, 508-509; weights in, "explicit," 506, in, fixed or fluctuating, 509, in, "implicit," 506, price and quantity, in, 509-510, relation of, to the purpose of, 507, sources of, for price, 488-489, value versus quantity, and, 505.
Index numbers, general business conditions, 537-543; Harvard

Committee on Economic Research, 538-541.
Index numbers, general price level, 544.
Index numbers, industrial stock prices, 545.
Index numbers, international prices, 544.
Index numbers, labor market, 544.
Index numbers, price and money rates, 544.
Index numbers, production: agriculture, 531-532; agriculture, mining, and manufacture, combined, 534-535; Department of Commerce's, 535; Federal Reserve Board's, 535; manufacture, 533-534; mining, 532-533.
Index numbers, retail prices; problems in calculating, 520; United States Bureau of Labor Statistics', of cost of living, 521-523, of foods, 520-521.
Index numbers, trade: Department of Commerce's, 537; Federal Reserve Board's, 537.
Index numbers, velocity of bank deposits, 544.
Index numbers, volume of trade: Persons', 536; Snyder's, 536-537.
Index numbers, wages of common labor, 546.
Index numbers, weekly earnings in New York State, 546.
Index numbers, wholesale prices: Annalist's, 527; Bradstreet's, 523-525; Dun's 525-526; Federal Reserve Board's, 518-519; Fisher's, 528-529; United States Bureau of Labor Statistics', 516-518.
Informants: method of securing good-will of, 53.
Interviews: method of collecting primary data by, 65. (*See* Collecting data, Collecting primary data.)

J-shaped distribution: form of 381.

Lag: use of, in correlating adjusted time series, 461-463. (*See* Correlation, Time series.)
Law of large numbers: nature of, 2.
Least squares: formula for slope of the line of, 445; illustration of the calculation of the line of, 445-447; method of, contrasted with moving average, 447, in measuring secular trend, 444-448. (*See* Lines of "best fit," Regression lines, Trend, Secular change.)
Linear relations: explanation of, 413.
Line of "best fit": extension of, in time series, 447; nature of, 447. (*See* Least squares, Regression lines.)
Lines: use of, to connect successive ordinates in continuous series, 227-229, in discrete series, 224. (*See* Diagrammatic presentation, Diagrams, Graphic presentation.)
Lines, regression (*See* Regression lines).

Mandatory power: use of, 52. (*See* Collecting data.)
Maps, statistical: choice of colors in, 203-204; contrast of, and tabulation, 201; functions of, 201-202; psychological bases for the use of, 201-202; types of, 203-212, dot, 207-212; uses of colored, 203-204, cross-hatched, 204-206, dot, 207-212, frequency dot, 210-212, shaded dot, 208-209, varying sized dot, 207-208.
Mean (*See* Arithmetic mean, Geometric mean).
Median: amount in time series, 290-294; arithmetic mean and, contrasted, 282; definition of, 264; deviations from, a minimum, 337-338; formula for determining the, with an even number of measurements, 283, with an odd number of measurements, 283; graphic location of the, 289, in cumulated time series, 290-294; interpolation for the, in discrete and in continuous series, 286-289; meaning of abscissa in graphically locating the, 290; methods of calculating the, 282, 283-294; method of de-

termining in a grouped series, 286, in an ungrouped series when *n* is even, 285-286, in an ungrouped series when *n* is odd, 283-285; partition expression as the, 282; period in time series, 290-294; properties of, 310-311; use of, in summarizing link-relatives, 450-455, as a type, in measuring dispersion, 337-338, in a graphic measurement of dispersion, 335; what it is, 282-283. (*See* Averages, Dispersion, Time series.)
Median chain relatives: methods of adjusting, in measuring seasonal variation, 451-452. (*See* Index numbers, general; Time series.)
Median link relative: method of, in measuring normal seasonal variation, 450-455. (*See* Index numbers, general; Time series.)
Median of relatives (*See* Index numbers, general).
Medians: use of moving, in measuring normal seasonal variation, 449-450. (*See* Time series.)
Mode: definition of the, 264, 294; different uses of the term, 294-295; graphic location of the, in cumulated frequency series, 305, in simple frequency series, 304, in space series, 300, in time series, 299-300; group adjustment in order to secure a, 305-306; the ideal, 295-296; interpolating for the, 382-384, in continuous series, 301-304; location of the, in discrete series, 301, in frequency series, 300-307, in space series, 300, in time series, 297-299; precision of the, 296-297; properties of the, 310-311; summary of the characteristics of the, 307; use of the, as a type, in measuring dispersion, 337-338, in summarizing link-relatives, 451; what it is, 294-297. (*See* Averages.)
Moving averages: method of calculating, 257-258.
Moving-range: in frequency series, 328; in historical series, 327-328. (*See* Dispersion.)

Normal distributions: characteristics of, 382. (*See* Normal law of error curve.)
Normal law of error curve: basis for expecting the, 399; chance phenomena and the, 399-402; description of the, 230; form of the, 368; general aspects of the, 360-375; illustration of the shape of the, 378; price dispersion and the, 493-495; properties of the, 367-370; relation of standard and average deviations in the, 351; relation of the standard deviation to total frequencies in the, 351-352. (*See* Probability distribution.)
Normal seasonal variation: use of averages in measuring, 449; use of link relatives in measuring, 450-455. (*See* Time series.)

Ordinate: definition of the, 219; relation of the, to the abscissa in plotting time series, 244-245; graphic representation and the, 221-222; time series and the units on the, 244-255. (*See* Graphic presentation.)

Pearsonian coefficient of correlation (*See* Correlation, Coefficient of correlation).
Pictograms: uses of, for illustrating discrete series, 175-198, as diagrams, 182-185, 193-195. (*See* Diagrammatic presentation, Diagrams.)
Point of origin: definition of the, in a system of co-ordinates, 219. (*See* Graphic presentation.)
Population: problems in counting, 57-59; use of statistics in a study of, 19.
Price dispersion: significance of, and index number making, 495-496. (*See* Dispersion; Index numbers, general.)
Price index numbers (*See* Index numbers, general; Index num-

bers, retail prices; Index numbers, wholesale prices.)
Prices: difficulties in securing, and index number making, 487-488; use of statistics in a study of, 19; variation of, and index number making, 486-487. (*See* Index numbers, general.)
Primary data: collecting, 47-71; conditions preliminary to the collection of, 46-53; contrast of, with secondary data, 24-26; definition of, 25; restrictions on the use of, 52; sanction for the collection of, 52-53; sources of, and the collection process, 61-64; types of problems requiring, 49-50; ways in which secured, 65-68.
Probability distribution: illustrations of, 361-364; tossing of ten coins and the, 366. (*See* Normal law of error curve.)
Probable error: formula for the, of the correlation coefficient, 429, of the mean, 370, standard deviation, 370; correlation coefficient and its, 429; correlation coefficient and the, in time series, 464-465; meaning of the, 370-372, in time series, 465; sample measurements and the, 372-373. (*See* Correlation, Time series.)
Production: use of statistics in a study of, 21.
Production index number (*See* Index numbers, production).
Profits: use of statistics in a study of, 19.

Quartile deviation: calculation of dispersion by the, 356-358; formula for the, measure of dispersion, 356; relation of the, to the standard deviation, 356, to the total frequencies, 356. (*See* Dispersion.)
Quartiles: coefficient of skewness and the, 385-386; formulæ for locating the, 289, 357; graphic location of the, 289; meaning of the, 289; meaning of abscissa scale units when, graphically located, 290; measure of skewness based on the, 385-386; not in the nature of averages, 289.
Questionnaires: editing of, 68-70; form of, 66-68, in a study of wages, 117-122; making out of, 56-57; rules for the use of, 66-67; use of, 66-68. (*See* Collecting primary data, Schedules.)

Range (*See* Dispersion).
Ratio changes: methods of showing, 249-252. (*See* Graphic presentation.)
Ratio charts: advantages of, 253-255; uses and merits of, 248-255. (*See* Graphic presentation.)
Ratio scale: zero on, 244.
Ratios: list of corrected, 89; relating to condition, 82-84, to space, illustrations of, 81, to time, illustrations of, 80-81; tests of good, 83. (*See* Units.)
Ratios of averages (*See* Index numbers).
Real wages: definition of, 96, 97.
Regression: formula for the slope of the line of, 445.
Regression coefficient: meaning of the, 426-428.
Regression lines: explanation and meaning of, 425-428; formulæ for the slope of, 426-427; lines of "best fit" as, 425-428; methods of drawing, 426-428.
Relatives, chain: method of calculating, 451-452. (*See* Index numbers, Time series.)
Relatives, link: use of arithmetic mean in summarizing, 451; use of median in summarizing, 450-455. (*See* Index numbers, Time series.)
Rents: use of statistics in a study of, 21.

Salaries: definition of, 96.
Salary-rates: definition of, 96.
Sample: test of a good, 32.
Samples: illustrations of the use of, 33-35; secondary data as, 32.
Sampling (*See* Collecting data, Collecting primary data).

Scale adjustment: method of, in difference charts, 246-248. (*See* Graphic presentation.)
Scale conversion: method of, in plotting time series on a difference basis, 246-248. (*See* Graphic presentation.)
Scale units: graphic presentation and the abscissa, 222, and the ordinate, 221-222; relation of, in difference charts, 244-247. (*See* Graphic presentation.)
Schedules: editing of, 68-70; form of, 66-68; form of, in a study of wages, 119-122; making out of, 56-67; rules for the use of, 66-67. (*See* Collecting data, Questionnaires.)
Seasonal variation: adjusted monthly indexes of, 453; explanation of, in time series, 440; measurement of, by monthly means or averages, 449-450; measurement of changing, 450; median-link-relative method of measuring, 450-455; methods of adjusting median chain relatives in measuring, 451-452; methods of measuring normal, 448-455; use of moving-median in measuring, 449-450. (*See* Correlation, Time series.)
Secondary data: accuracy of, 38-42; as samples, 32; bias and, 32; characteristics of, 24-25, 45-46; contrast of, with primary data, 24-26; definition of, 24; distinguished from primary data, 26; germane to a problem, 44-46; homogeneity of, 42-44; nature of, 31-35; purpose for which issued, 31; sources of, 26-31; sources of weakness of, 34-35; tests to be applied to, 30-46; types of consumers of, 31; units in which, expressed, 35-38.
Secular change: explanation of, 440; trend as a generalization of, 440; free-hand method of measuring, 443-444; least-square method of measuring, 444-448; meaning of, 442-443; measurement of, by method of averaging, 444; method of eliminating, 447; methods of measuring, 442-448. (*See* Time series.)
Series: arbitrary nature of grouping in continuous, 216; characteristics of condition, 158-169, of continuous frequency, 162, 166-167, of discrete frequency, 162-164, of historical, 157, of space, 157-158; continuous as to unit and discrete as to number, 215-216; definition of continuous frequency, 162, of discrete, 162, of a statistical, 157; desirable things to know about statistical, 325; diagrams unsuited to illustrate continuous, 215-216; distribution of items in continuous, 166-167, in discrete, 163-164; general treatment of time, 438-467; graphic presentation of cumulated continuous frequency, 237-242, of cumulated discrete frequency, 233-237, of cumulated frequency, 232-242, of frequency, 221-242, of simple frequency, 221-232; grouping of data in continuous, 231-232; illustrations of continuous, 214-215, both as to unit and measurement, 217-218, space, 218; illustrations of discrete, 214-215; location of the mode in continuous, 301-304, in discrete, 301; lines connecting successive ordinates and discrete, 223-227; measurements in either discrete or continuous, 214; methods of analyzing statistical, 393-394; nature of continuous, 214-215, of measurements in discrete, 222; order of arrangement in space, 157-158; plotting frequency distributions describing continuous, 227-232, discrete, 222-227; plotting simple historical, 243-258; smoothed lines connecting successive ordinates and continuous, 227-229, and discrete, 225-227; smoothing continuous, frequency, 229-232; time, space, and condition, contrasted, 157-169; treatment of lines connecting successive ordinates in cumulated continuous frequency,

239-242; types of statistical, 157, and corresponding tables, 157-169; units on the abscissa, and continuous frequency, 227. (*See* Space series, Time series.)
Simple unit: definition of a, 35-36, 77-78. (*See* Units.)
Skewed distributions: characteristics of, 379; form of ideal moderately, 379; types of, 378-381. (*See* Skewness.)
Skewness: cause of, in distributions, 378; coefficients of, 381-386, based on extremes, median, and quartiles, 386, based on mean and mode, 384-385; definition of, 384; dispersion contrasted with, 377-378; function of measures of, 377; general aspects of, 376-392; measure of, 381-386, based on mean and mode, 382-385, based on the quartiles, 385-386; test of negative, 381, of positive, 381; typical of distributions, 376-378.
Smoothed frequency curve: illustration of a, 231. (*See* Graphic presentation.)
Smoothing: continuous frequency series and curve, 229-232; time series and curve, 256-258; use of curve, 225-227. (*See* Graphic presentation.)
Space series, illustration of continuous, both as to unit and measurement, 218. (*See* Series.)
Standard deviation: calculation of the, in frequency series, 354-355, in time series, 352-353, in time series from an assumed average, 353, in time series from the true average, 352; coefficient of, 355; coefficient of skewness and the, 384; effect of extreme deviations on the, in historical and in frequency series, 349-350; formula for the, 349; measure of dispersion based on the, 349-358; method of calculating the, 349; relation of the, to the average deviation, 351, 355, to the total frequencies, 351-352; sum of the squares a minimum in, when deviations are taken from the arithmetic mean, 350. (*See* Dispersion.)
Standard Statistics Corporation (*See* Index numbers, barometric and forecasting).
Standards for collection of data, 22.
Statistical data: characteristics of a body of, 125-126; characteristics of, expressed in a series, 126; cumulative characteristics of, 126; estimates as, 60; types of, and test for their use, 22-46; ways of collecting primary, 65-68.
Statistical data not a matter of record: collection of, 57-59.
Statistical methods: abuse of, 5-6; application of, to business units, 15, to economic theory, 17-18, to groups of business units, 15-16, not universal, 4, to public affairs, 16-17, to social economy, 16; definition of, 12; limitations of, 14; progress in the use of, 13; use and application of, 13-21.
Statistical series: characteristics of condition, 158-169, of historical, 157, of space, 157-158; desirable things to know about, 325; types of, 157, and corresponding tables, 157-169. (*See* Series.)
Statistical units: diagrammatic illustration of different types of, 88. (*See* Units.)
Statistics: as methods, 12; characteristics of, 45-46; contrasted with accounting, 1-2, with statistical methods, 9; definitions of, 9-10, explained, 10-12; derivative nature of, 24; fundamentals of, as methods, 77; meaning and application of, 1-21; method of collecting, 54-62; nature of misuse of, by beginners, 6-8; points in identification of, 24; rules in use of, 6; scope of, 3; source of, and the collection process, 61-64, as finished products, 22-23, as raw material, 22-23; steps in the use of, 5; synthetic nature of, 23-24; use of, as "proof," 4, in study of the business cycle, 18-19, in study of

competition, 19, in study of consumption, 19, in study of costs, 19, in study of demand, 19, in study of population, 19, in study of prices, 19, in study of production, 20-21, in study of profits, 19, in study of rents, 21, in study of trade, 21, in study of unemployment, 21, in study of wages, 21, in study of wealth and income, 19. (*See* Statistical methods.)

Stub: contents of, in tables, 128-129. (*See* Tables, Tabulation.)

Symmetrical distribution: forms of an ideal, 378. (*See* Normal law of error curve.)

Table: illustration of a "four-part" double frequency, 431.

Table structure: rules governing, 143-144.

Tables: contents of, 146-148; contents of captions in, 128-129, of the stub in, 128-129; derivative, 139; emphatic parts of, 135; general characteristics of, 138-139; rules for construction of, 146-148; illustrations of faulty titles of, 149-151, of frequency, 160, 161, 164, 165, varying orders of items in, 136; inelastic character of, 129; items in, arranged according to alphabet, 135, to size or frequency, 132-133, to space, 133-134, to time, 133, to a variable condition, 134; items in, ranked, 132-133; nature of frequency, 158-169; necessity of logical arrangement of items in, 135; numbering lines and columns in, 146; order of the arrangement of data in, 132-134; positions of totals in, 144-145; rules for construction of general, 146-147, of summary, 147; size of, 145; titles of, 148-151; treatment of groups in frequency tables, 167-169; types of statistical, 138-139. (*See* Tabulation.)

Tabular arrangement: advantages of, 131-138.

Tabulation: advantages of the card system of, 155; characteristics of data and, 124-126; characteristics of hand cards for, 152-153, of machine cards for, 153; coding of details for, 151-152; contrast of, with diagrammatic presentation, 171-173; definition of, 124; meaning of, 128-131; mechanics of, 151-156; points of view in, 124; purposes of, 129; relation of, to classification, 128-131. (*See* Tables.)

Tabulation card: advantages of a machine, 154-155; illustration of a hand, 153, of a machine, 154.

Tabulation form: complexity of, 143; "double" type of, 140-141; "quadruple" type of, 142-143; "single" type of, 140; "treble" type of, 141.

Tabulation forms: classes of, 139-143.

Tests to be applied to secondary data, 30-46. (*See* Secondary data.)

Theory of error: binomial expansion and the, 363-367. (*See* Normal law of error curve.)

Theory of probability: general, 360-375. (*See* Normal law of error curve.)

Time changes: types of, 257. (*See* Time series.)

Time series: accidental changes in, explained, 441; calculation of the average deviation in, 339-342; characteristics of items in, 457; choice and adjustment of scales in plotting, 243-255; correlation of, 457-464; cyclical changes in, explained, 440; freehand method of showing long-time changes in, 439; general treatment of, 438-467; graphic presentation of, 242-259; illustration of continuous, both as to unit and measurement, 217-218; lines connecting successive ordinates in plotting, 255-256; location of median amount in, 290-294, period in, 290-294; meaning of abscissa in, 256; methods of measuring and isolating changes in, 441-457; na-

ture of changes in, 438-441; order of items in, 464; plotting cumulative, 258-259, simple, 243-258; probable error of the correlation coefficient in, 464-465; problems in graphic presentation of, 243; purposes and methods of smoothing, 256-258; random sampling and, 464; seasonal variation in, explained, 440; short-time changes in, classified, 440; treatment of, conclusion, 465-466; use of the correlation coefficient in adjusted, 461-463, in unadjusted, 461-463. (*See* Correlation.)
Trade: use of statistics in the study of, 21.
Trend: determination of, by method of least squares, 445-447; free-hand method of measuring secular, 443-444; illustration of, in pig-iron production, 442; least-square method of measuring secular, 444-448; meaning of, in time series, 442-443; measurement of secular, by method of averaging, 444; method of determining monthly increment of, by least squares, 445-447; method of eliminating secular, 447; methods of measuring secular, 442-448; time series and secular trend, 440.

U-shaped distribution: form of, 380.
Unemployment: use of statistics in a study of, 21.
Unit: definition of a composite, 36, 78, a simple, 35-36, 77-78.
Units: coefficients as, defined, 36-37; meaning of statistical, of measurement, 72-77; meaning of abscissa, in discrete series, 222-223; types of, 35-38.
Units of analysis: diagrammatic illustration of, 88.
Units of analysis and of interpretation, 80-84.
Units of enumeration: diagrammatic illustration of, 88.
Units of enumeration or estimation, 77-80.
Units of measurement: classification of, 77-84; general, 72-80; rules for use of statistical, 90.
Units of presentation: classified, 85; diagrammatic illustration of, 88; nature of, 85-88; rules for use of statistical, 90.
Units of presentation involving condition: test of crudity of, 86.
Units of presentation involving space: test of crudity of, 85-86.
Units of presentation involving time: test of crudity of, 85.

Wage data: employes as sources of primary, 101, of secondary, 104-106; employers as sources of primary, 101-102, of secondary, 106-109; sources of secondary, 104-116; trade and labor unions as sources of primary, 102-103, of secondary, 109-115.
Wage-rate as a statistical unit, 92-93.
Wage-rates: definition of, 96.
Wages: bases for a definition of, 95-96; confusion in the use of the term, 94-95; declaration of the purpose of a study of, 117-118; definition of, 96, in a study of, 117, and use of terms, 97-100; primary data on, in relation to specific problems, 100-104; study of, to illustrate statistical methods, 92-123; use of statistics in the study of, 21.
Wealth: use of statistics in the study of, 19.
Weighted averages: simple and, contrasted, 279-281. (*See* Index numbers, general.)
Weighting (*See* Index numbers, general).
Weights: sources of, for index numbers of prices, 488-489.

Zero: abscissa and, position, 221; difference scale and, 244; horizontal, in time or historical series, 243; ratio scale and vertical, 244.
Zero cases included in averages, 281.

# PERSONAL INDEX

(The numbers refer to pages and include citations in both text treatment and chapter references.)


Adamson, Tilden, 80.
Averill, W. A., 473.

Bailey, W. B., 71, 213.
Barnett, George E., 513.
Berridge, W. A., 21, 123, 544.
Bivins, P. A., 251, 252, 259.
Black, S. B., 398.
Bliss, G. I., 316.
Boerner, E. G., 24.
Booth, Charles, 105.
Boucke, O. Fred., 21.
Bowley, A. L., 8, 10, 12, 20, 44, 46, 64, 71, 72, 87, 91, 170, 213, 245, 260, 322, 338, 358, 374, 391, 407, 410, 412, 436, 459, 508, 513.
Brinton, W. C., 212, 213, 260.
Brown, William, 406.
Burgess, W. Randolph, 544, 546.
Burnett-Hurst, A. R., 64.

Carr, Elma, 522.
Chapin, F. Stuart, 46, 71, 123.
Chapin, R. C., 105.
Clark, Earle, 221, 349, 358.
Coffey, P., 23, 318.
Coyle, Eunice S., 529, 530.
Cramer, Frank, 87, 127.
Crathorne, A. R., 437.
Crum, W. L., 466.
Cummings, John, 71.

Darbishire, A. D., 400.
Davenport, C. B., 406.
Davenport, Eugene, 228, 232, 358, 373, 399, 406, 436.
Davies, G. R., 436, 449, 466, 513.
Davis, J. S., 545.
Day, E. E., 20, 170, 173, 531, 532, 533, 534, 535.
Dewey, Davis R., 107, 288.
Dublin, Louis I., 380.
Durand, E. D., 170.

Elderton, E. M., 260, 322, 358, 436
Elderton, W. P., 260, 322, 358, 407, 436.

Falkner, H. D., 448, 466.
Falkner, Roland P., 115, 199.
Field, James A., 252.
Fisher, Arne, 360, 373, 374.
Fisher, Irving, 19, 20, 255, 260, 322, 482, 483, 491, 509, 510, 511, 513, 528, 529.
Flux, A. W., 514.
Forsyth, C. H., 436.
Foster, W. T., 19, 467.
Frankfurter, Felix, 398.
Frickey, Edwin, 447, 545.

Galton, Sir Francis, 406, 425.
Gide, Charles, 95.
Goldenweiser, E. A., 545.

Hall, L. W., 466.
Haney, L. H., 29.
Hansen, A. H., 18.
Harris, J. Arthur, 406.
Hart, Hornell, 544.
Haskell, Allan C., 213, 260.
Hayford, John F., 266.
Hedrick, Earle R., 547.
Hettinger, A. J., 19, 467.
Hight, J., 5.
Holdaway, C. W., 397.
Hooker, R. H., 395, 406, 407, 514.

Insull, Samuel, 413.

Jevons, W. Stanley, 17, 170, 360, 365, 367.
Johnson, A. S., 95.

Jones, D. Caradog, 20, 262, 322, 358, 360, 374, 391, 436.
Jones, Richard, 17.

Karsten, Karl G., 185, 213, 260.
Kelley, Truman L., 260, 280, 322, 359, 374, 436.
Kemmerer, E. W., 20.
Kent, F. C., 374.
Keynes, J. M., 20, 262, 360, 374.
Keynes, J. N., 18.
Kincer, J. B., 436.
King, W. I., 20, 21, 46, 123, 170, 213, 232, 260, 322, 359, 392, 396, 436, 444, 448, 449, 450, 466, 531.
Knauth, Oswald W., 543.

Leonard, W. E., 21.
Lindeman, Eduard C., 261.

Macaulay, F. R., 514.
Marshall, Alfred, 260.
Marshall, W. C., 213, 260.
Matthews, A. M., 545.
Maxwell, F. W., 545.
McIlraith, James W., 5.
Meeker, Royal, 43, 514.
Meyer, C. C., 29.
Mills, Frederick C., 12, 20, 260, 322, 359, 375, 436, 467, 514.
Mitchell, W. C., 18, 20, 330, 331, 334, 359, 392, 397, 473, 479, 480, 482, 484, 485, 486, 487, 488, 489, 491, 493, 495, 496, 499, 500, 501, 506, 507, 509, 511, 514, 545.
Moore, H. L., 18, 21, 407, 437.
More, L. B., 105.
Mowbray, A. H., 398.
Mudgett, B. D., 29.

Nearing, Scott, 95, 104, 116, 123.
Newsholme, Arthur, 6.

Ogburn, W. F., 19.

Parmelee, Julius H., 29.
Patton, A. C., 466.
Pearl, Raymond, 10, 14, 19, 20, 220, 359, 360, 370, 371, 372, 437.
Pearson, Karl, 20, 399, 406, 407, 410, 437.
Persons, C. E., 98.
Persons, W. M., 18, 19, 21, 39, 407, 408, 437, 441, 443, 444, 447, 448, 450, 455, 457, 463, 464, 465, 467, 511, 514, 529, 530, 536, 538, 541, 544.

Reed, L. J., 19.
Rietz, H. L., 228, 229, 232, 359, 398, 437, 463.
Ripley, W. Z., 203.
Roberts, Elmer, 398.
Rowntree, B. Seebohm, 105.
Rubinow, I. M., 317, 318.
Rugg, H. O., 20, 46, 71, 123, 323, 359, 375, 428, 437.
Rutter, Frank R., 43.

Schultz, Henry, 19.
Seager, H. R., 95.
Secrist, Horace, 19, 20, 21, 279, 281, 296, 305, 315, 323, 391.
Sigwart, C., 319, 320.
Simpson, Kemper, 19.
Slichter, C. S., 367, 375.
Snider, Joseph L., 514
Snodgrass, Katherine, 544.
Snyder, Carl, 21, 514, 531, 536, 538, 544.
Stewart, W. W., 21, 531.
Streightoff, F. H., 104, 116, 123.
Strong, T. B., 400.

Taussig, F. W., 19, 469.
Thorndike, E. L., 260, 359, 406.
Tolley, Howard R., 377.

Vance, Ray, 541.
Venn, John, 170, 263, 322, 323.
Vernon, H. M., 406.

Watkins, G. P., 91, 170, 263.
Weldon, W. F. R., 400, 401.
West, Carl J., 20, 29.
Whipple, G. C., 375.
Whipple, Guy Montrose, 373, 406, 437.
Wright, P. G., 19.
Wright, T. W., 266.

Young, A. A., 514.
Yule, G. Udny, 10, 20, 260, 266, 323, 350, 351, 352, 356, 359, 360, 378, 379, 392, 407, 410, 425, 437.

Žižek, Franz, 91, 170, 288, 323, 359.